Titre: Title:	Analyse logique de données pour estimer le taux de présence des passagers en transport aérien.
Auteur: Author:	Christine Dupuis
Date:	2010
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dupuis, C. (2010). Analyse logique de données pour estimer le taux de présence des passagers en transport aérien. [Master's thesis, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/283/

UNIVERSITÉ DE MONTRÉAL

ANALYSE LOGIQUE DE DONNÉES POUR ESTIMER LE TAUX DE PRÉSENCE DES PASSAGERS EN TRANSPORT AÉRIEN

CHRISTINE DUPUIS

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(MATHÉMATIQUES APPLIQUÉES)

AVRIL 2010

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

ANALYSE LOGIQUE DE DONNÉES POUR ESTIMER LE TAUX DE PRÉSENCE DES PASSAGERS EN TRANSPORT AÉRIEN

Présenté par : DUPUIS, Christine

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

M. AGARD Bruno, Doct., président

M. GAMACHE Michel, Ph.D, membre et directeur de recherche

M. BIGRAS Louis-Philippe, Ph.D, membre

REMERCIEMENTS

En premier lieu, je tiens à remercier très sincèrement mon directeur de recherche, le professeur Michel Gamache, pour son excellent soutien depuis le tout début jusqu'à la fin du projet. Ses conseils constructifs, sa disponibilité et son optimisme m'ont motivés à conduire mon projet jusqu'au bout. Malgré toutes les épreuves rencontrées et les moments plus difficiles, il s'est toujours montré très compréhensif. Je le remercie donc de tout cœur d'avoir eu confiance en moi et de m'avoir encouragée à persévérer.

Je me dois également de remercier Air Canada pour toutes les ressources qui ont été mises à ma disposition et sans lesquelles ce projet n'aurait pu voir le jour. Plus particulièrement, merci à Jean-François Pagé, chef de service de l'équipe de recherche opérationnelle, de m'avoir encadrée pour la partie industrielle du projet, mais aussi pour ses conseils judicieux et sa très grande générosité. Je tiens également à saluer mes collègues d'Air Canada: Jérôme, Jacques, Karine, Julie, Soad, Peter, et tous les autres, qui ont partagé mon quotidien durant ces deux dernières années et qui n'ont jamais hésité à m'offrir leur aide. J'ai beaucoup apprécié la dynamique chaleureuse de cette équipe, et ce fut un véritable plaisir de travailler à leurs côtés.

Je ne pourrais passer sous silence la contribution financière du Conseil de recherche en sciences et génie (CRSNG), envers qui je souhaite exprimer ma gratitude. Merci aussi à l'École Polytechnique de Montréal, de laquelle je conserverai d'excellents souvenirs tant du baccalauréat que de la maîtrise. Je voudrais aussi remercier Suzanne Guindon, ainsi que tout le personnel enseignant et non-enseignant du Département de génie industriel et de mathématiques.

Je remercie également Alexe Sorin de m'avoir accordé la licence de son logiciel et d'avoir répondu à mes questions quant à son utilisation, sans quoi ce projet n'aurait certainement pas été possible.

Finalement, durant ma maîtrise, j'ai eu la chance de présider un exécutif absolument extraordinaire à l'AÉCSP: Marc-Alexandre, aussi mon voisin de cubicule chez Air Canada, Maléna et Amélie, d'excellentes amies, Amandine et Joël, qui ont été aussi mes colocataires, et finalement, Simona et Jonathan. Ces personnes m'ont inspirée, elles m'ont encouragée et surtout, elles ont cru en moi. Aujourd'hui, je tiens à leur exprimer ma plus profonde reconnaissance, et aussi leur réitérer à quel point c'est pour moi un privilège de les connaître.

Mes derniers remerciements vont aux deux personnes qui m'ont inconditionnellement soutenue à tous les niveaux: mes parents, Sylvie et Jean-Guy Dupuis.

RÉSUMÉ

Chaque année, dans l'industrie du transport aérien, des pertes de revenus additionnels estimées à des millions de dollars sont causées par des passagers absents. En effet, ces sièges qui ont été vendus mais qui seront inoccupés peuvent potentiellement être revendus à d'autres passagers si on est capable d'en estimer le nombre correctement. Cela génère des profits supplémentaires pour les compagnies aériennes, à condition de ne pas sur-utiliser cette façon de faire, car un passager à qui l'on refuse l'embarquement dû à un manque de place sur l'avion devient coûteux, puisqu'il faut le dédommager.

Le projet de maîtrise consiste en l'élaboration d'un modèle permettant de mieux prévoir le nombre de sièges supplémentaires par rapport à la capacité initiale de la cabine que l'on peut se permettre de vendre, phénomène appelé la survente. L'approche retenue est le « Logical Analysis of Data », auquel nous ferons référence par la méthode LAD. Plus spécifiquement, le modèle classifie les passagers en trois groupes: présents, absents et incertains, chaque groupe possédant son propre taux de présence. La somme pondérée de ces trois groupes et de leurs taux respectifs constitue le nombre de personnes présentes prévues par la méthode LAD.

Cette méthode a été retenue à cause de son originalité et de ses succès connus à ce jour. Elle se distingue des autres formes de data mining plus conventionnelles par le fait qu'elle fait preuve d'une certaine forme d'intelligence artificielle; à partir des caractéristiques des passagers, elle établit des combinaisons de conditions (appelées patrons) pour lesquels les passagers ciblés ont une plus forte tendance à être présents (ou absents). Les caractéristiques sont par exemple la classe de réservation, le jour de la semaine du départ, l'heure, l'origine de l'itinéraire…

En bref, la méthode LAD comporte quatre grandes phases. La première consiste à indexer les données, ou du moins, les rendre numériques. En effet, la méthode LAD ne fonctionne pas avec des attributs catégoriques. La deuxième phase est la discrétisation de ces données à l'aide d'une grille que l'on appelle système de points de coupures. Les données sont ensuite renumérotées en fonction de l'intervalle de la grille à l'intérieur duquel elles se situent.

La troisième phase est la génération des patrons. La première moitié est lue par un algorithme aux termes duquel une liste de patrons est proposée. Un patron comporte un certain nombre de conditions ou d'attributs bornés; les bornes étant limitées aux points de coupures. L'utilisateur peut fixer certains paramètres afin d'encadrer la création de ces patrons: l'homogénéité (la proportion de personnes présentes à l'intérieur du groupe), la prévalence (la proportion de données inclues dans ce patron par rapport aux données totales du même type) et le degré (le nombre de conditions maximum permises dans un patron). La deuxième moitié des données est ensuite évaluée en fonction des patrons créés et retenus, et classifiée à son tour en trois groupes.

La quatrième phase est la sélection des patrons parmi ceux générés. Les patrons retenus constituent le modèle. Les taux de présence de chacun des groupes (positifs, négatifs et incertains) sont égaux aux homogénéités de ces mêmes groupes.

Pour réaliser l'ensemble de ce travail, un logiciel d'Alexe Sorin, le LAD Datascope V2.0 développé au RUTCOR Center de Rutger's University du New Jersey, a été utilisé. Bien que les deux premières phases puissent être prises en charge par le logiciel, nous avons choisi de faire ces étapes nous-mêmes, à l'aide d'Excel et d'Access. Ensuite, le logiciel a été utilisé pour la génération des patrons et l'évaluation de ceux-ci. Afin de comparer les résultats avec le système de prévisions actuellement en place chez Air Canada, nous

avons également eu recours à plusieurs programmes en Visual Basic pour lire les patrons et classifier les observations.

Lorsque mis en comparaison avec PROS, l'outil de prévision utilisé actuellement chez Air Canada et basé sur un historique, la méthode LAD se montre très concurrentielle. Effectivement, la somme des erreurs au carré par rapport aux présences réelles est réduite de la moitié et l'écart type aux deux tiers. De plus, lorsque l'on observe les coefficients de relation entre le nombre de personnes réelles et la méthode LAD, on obtient un excellent R de 0,99635, tandis que PROS n'obtient que 0,97118. Ceci signifie que la méthode LAD démontre une corrélation plus forte avec les observations actuelles que PROS.

Éventuellement, l'application de cette méthode pourrait être étendue à d'autres paires de villes, mais d'abord la méthode nécessite encore d'être raffinée et exploitée à son plein potentiel. Il y a plusieurs étapes dans la méthode LAD que nous n'avons pas encore explorées en profondeur comme les systèmes de points de coupures ou le choix des patrons à retenir pour le modèle parmi tous ceux générés.

Cette méthode peut également être utilisée à d'autres fins en transport aérien. Une suggestion consiste à mettre en place une méthode LAD qui ne classifierait pas les passagers, mais plutôt les vols. En effet, les vols pourraient être classifiés en plusieurs tranches selon le taux de présence prévu. Puisque chaque observation (vol) possède son propre score, on pourrait utiliser plusieurs valeurs critiques pour répartir les observations en groupes ayant des taux de présence similaires. Pour illustrer ceci, le premier groupe de vols aurait par exemple des taux de présence très faible, et le dernier groupe les taux de présence les plus élevés. Lorsque l'on calculerait le score pour un vol, on regarderait entre quelles valeurs critiques il se trouve pour établir son taux de présence.

ABSTRACT

In the airline industry, revenue losses are estimated to reach millions of dollars yearly due to passengers that don't show up for their flights, this is referred to as «no-shows». A frequent practice in the airline industry is to overbook flights to make up for these losses. Some significant revenues can be generated by this practice if the forecasts are accurate. If the no-show forecast is too low, potential revenue loss will remain. On the other hand, if the forecast suggests too many no-shows, some passengers may be denied boarding. This has a direct negative impact on customer satisfaction, and it is difficult to determine the exact cost of customer's frustration.

The objective of this master's project is to build a model that would improve the accuracy of predictions for show and no-show passengers, and consequently adjust the overbooking levels. The chosen method is known as the «Logical Analysis of Data», also referred to as LAD. Specifically, this method classifies all passengers into three groups: positive (showing up), negative (no-shows) and unclassified. Each of these three groups has its own show rate. The weighted sum of these groups and their show rate results in the total show rate for the evaluated group of passengers.

This approach was chosen not only for its originality, but also for its success in various sectors. It differs from other conventional data mining methods by its ability to detect combinatory information about the passengers. The input consists of a number of observations (passengers), each described by a vector of attributes derived from characteristics such as booking class, day of the week, departure time, itinerary origin, … The LAD method detects sets of conditions on attributes for which the group of passengers respecting these conditions have a significantly higher or lower show rate.

The LAD method can be broken down into four phases. The first is to gather the data, and ensure that it's all numerical. LAD method does not work on categorical attributes. The second phase is to build a system of cutpoints or a grid of separation. The data is then indexed to fit its position in the grid.

The third phase is the pattern generation. After scanning the first half of the database chosen randomly, the LAD method proposes a list of patterns to the user. A pattern contains a number of conditions or bounded attributes: the choice of boundaries is limited to the cutpoints. The user must fix different parameters in order to guide the pattern generation: homogeneity (proportion of attending passengers in the group), prevalence (proportion of passengers included in the group versus all the passengers of this type), degree (maximum number of conditions that can be used for one pattern). The second half of the database is then evaluated in function of the pattern list, and classified into the three groups.

The last phase of this method is to select from the list of patterns those that will be included in the model; i.e. the final list of patterns or theory. For each of the three groups, the ultimate show rate is in fact equal to the homogeneity of the group. These rates are then applied to new passengers according to their classification.

In order to implement the whole method, the LAD Datascope V2.0, developed by Alexe Sorin, at RUTCOR Center, Rutgers's University in New Jersey, was used. Although both the first two phases can be performed with the software, we chose to develop our own tools, using mostly Excel and Access. The software has then been very helpful for the pattern generation and the evaluation. We also had to develop a few Visual Basic programs to be able to make the comparisons with Air Canada's actual methods of forecast and to classify new data. These programs can per example read the patterns, and classify the data.

Air Canada's tool for overbooking forecasts, PROS, is based on historical statistics for the flight. When compared to PROS, the LAD method appears to be very competitive. In fact, the sum of squared errors between actual observations and LAD method is only half of the one obtained by PROS. In addition, the standard deviation is reduced to two thirds of PROS value. Examination of the R coefficients of correlation also shows that the LAD method seems superior. The correlation between LAD predictions and actual observations is as high as 0,99635, while PROS coefficient is only 0,97118.

To further reap the benefit of this study, it is strongly recommended to spread the application of this method to other routes (city pairs), but first the method needs to be refined. It is also not used up to its full potential. There are still a few steps of the LAD method that we did not analyze deeply, such as cutpoints systems or selection of patterns from the list to build the model. Another suggestion consists in developing a new LAD model that would not classify the passengers as shows or o-shows, but flights instead. They could be classified into different groups according to their show rate.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

AGIFORS	Airline Group of the International Federation of Operational Research Societies
BO	Business object
LAD	Logical analysis of data
PNR	Passenger Name Record
TRA	Training
TES	Testing

LISTE DES ANNEXES

INTRODUCTION

Chaque année, dans l'industrie du transport aérien, des pertes de revenus additionnels estimées à des millions de dollars sont causées par des passagers absents. En effet, ces sièges qui ont été vendus mais qui seront inoccupés peuvent potentiellement être revendus à d'autres passagers si on arrive à en estimer le nombre correctement. Cela génère des profits supplémentaires pour les compagnies aériennes, à condition de ne pas sur-utiliser cette façon de faire, car un passager à qui l'on refuse l'embarquement dû à un manque de place sur l'avion devient coûteux, puisqu'il doit être dédommagé. De plus, cette situation affecte l'image de l'entreprise. Chez la compagnie à l'étude, Air Canada, cette pratique est présentement sous-utilisée; alors que par les années passées on pouvait faire jusqu'à 10% de survente, le niveau se situe actuellement entre 0% et 5 %. Lors des meilleures années, jusqu'à 400 millions de revenus additionnels étaient générés par la survente. Cette pratique a donc clairement le pouvoir de faire la différence entre une année profitable ou non.

Le projet de maîtrise consiste à développer une approche qui permettrait de mieux prévoir le nombre de sièges supplémentaires par rapport à la capacité initiale de la cabine que l'on peut se permettre de vendre, phénomène appelé survente. Il faut donc identifier correctement les passagers qui sont à risque élevé d'être absents et ceux qui ont de fortes chances d'être présents lors du décollage, afin d'estimer correctement le nombre de sièges qui peuvent être disponibles pour la survente.

Chaque passager est caractérisé par un vecteur d'attributs basés sur des caractéristiques pouvant être extraites des bases de données, telles que la classe de réservation, l'heure et le jour du départ, l'origine, la destination, etc. La stratégie consiste à déterminer un ou des sous-ensembles parmi ces attributs qui permettraient de classifier les passagers d'Air Canada comme étant présents ou absents sur le vol étudié. La construction du modèle de

classification se fait à l'aide des données recueillies une fois les vols partis, lorsque l'on connaît le statut final du passager.

L'approche retenue est une forme de data mining supervisé appelée analyse logique de données, mieux connu sous l'appellation anglaise *Logical Analysis of Data* (LAD). Pour la suite de ce mémoire, nous ferons référence à cette approche comme étant la méthode LAD. Cette méthode consiste à lire l'ensemble des observations ainsi que leurs caractéristiques et d'établir des ensembles de conditions récurrentes pour lesquelles les observations qui satisfont ces conditions ont la même finalité. On appelle ces ensembles de conditions des patrons. Lorsqu'un ou plusieurs ensembles de conditions décrivant des passagers qui ont été présents lors du départ d'un vol sont satisfaites pour un nouveau passager, on classe ce dernier comme présent ou positif. À l'opposé, si ce passager correspond plutôt aux patrons négatifs, il est alors classifié comme un passager à risque d'être absent ou encore comme une observation de classe négative. Par la suite, l'application de ce modèle sur de nouvelles données sera mise en comparaison avec le système actuel de prévision des ventes utilisé chez Air Canada (PROS) pour mesurer si des profits additionnels peuvent potentiellement être générés, tout en maintenant le taux de refus d'embarquement très bas. Ce taux est présentement à 3 refus par 10 000 passagers, mais Air Canada pourrait se rendre jusqu'à 15.

La méthode LAD est relativement récente et elle a été utilisée pour plusieurs problèmes de classification, notamment pour des diagnostiques médicaux. Par exemple, on évaluait si le patient satisfaisait les patrons positifs ou négatifs pour diagnostiquer s'il était atteint du cancer ou non. Le but de ce mémoire est donc d'appliquer la méthode LAD à un ensemble de voyageurs chez Air Canada, et de déterminer si cette approche peut être utilisée efficacement pour détecter les passagers qui ont une probabilité plus élevée de se présenter pour leur vol.

Le présent mémoire se divise en quatre grands chapitres : dans le premier, la méthode LAD est expliquée brièvement; on fait ensuite une révision sommaire de la littérature existante au sujet de la méthode LAD à ce jour, mais aussi sur les processus de survente employés présentement par les compagnies aériennes. Au deuxième chapitre, la méthode LAD telle que nous l'avons utilisée est décrite plus en détails, ainsi que la manière dont elle peut s'appliquer à la survente de sièges. Au troisième chapitre, on décrit les différentes tentatives d'obtenir des résultats, et finalement, au dernier chapitre, les résultats sont discutés et mis en comparaison avec le système actuel de prédiction des ventes en matière d'«overbooking». En conclusion, le lecteur trouvera des suggestions de pistes à explorer dans le but d'améliorer les résultats ou de réaliser d'autres avancées dans le domaine.

CHAPITRE 1 REVUE DE LITTÉRATURE

Il existe des méthodes de data mining traditionnelles bien connues qui peuvent être utilisées pour résoudre les problèmes de classification ou obtenir de l'information cachée, notamment la segmentation (*clustering*), les arbres de décision, ou encore, les règles d'association. Ces méthodes plus conventionnelles sont basées essentiellement sur des outils statistiques et sont brièvement survolées dans la première sous-section. La méthode LAD diffère des précédentes par sa nature logique ou booléenne ainsi que son caractère combinatoire. Au niveau de la littérature existante à ce jour, plusieurs articles ont été trouvés au sujet de la méthode LAD. Il y a d'abord tous ceux qui traitent de l'aspect mathématique de la méthode LAD; cet ensemble d'articles décrit la méthode dans ses moindres détails et est présenté dans la deuxième sous-section de ce chapitre. On y trouve, entres autres, plusieurs façons de discrétiser les données et de multiples algorithmes pour la génération des patrons. Dans la troisième section, on retrouve des exemples d'applications de la méthode LAD et des comparaisons entre les résultats obtenus par des méthodes de classification conventionnelles en data mining et ceux obtenus en utilisant la méthode LAD. La dernière sous-section porte sur la manière dont les compagnies aériennes gèrent la survente des sièges sur leurs vols actuellement.

1.1 Méthodes de data mining conventionnelles

La segmentation consiste à séparer un ensemble de données en plusieurs groupes homogènes selon certains critères, par exemple en fonction d'une mesure appelée la distance. Les distances entre deux observations sont généralement calculées comme la somme des différences au carré de tous les attributs; et on peut choisir d'accorder plus de poids à l'un ou l'autre des attributs. Le principe est la création d'un certain nombre - généralement choisi par l'utilisateur - de groupes, les plus différents possibles les uns par rapport aux autres (grande distance), mais le plus homogène possible à l'intérieur (courte distance). Il existe plusieurs algorithmes de segmentation : Kohonem's SOM, K-means,

Diana, SVM, CURE, etc. Les différences principales résident dans le choix des points de départs et la formule du calcul de la distance. Les groupes formés constituent le résultat final. Cette méthode appliquée aux passagers aurait pour effet de répartir les passagers en plusieurs groupes distincts (types), pour lesquels nous pourrions observer des taux de présence ou d'absence significativement différents d'un groupe à l'autre. Sachant que les passagers d'un même groupe se ressemblent, les attributs qui caractérisent chaque groupe pourraient alors être utilisés pour décrire chacun des types de passagers. En classifiant de nouveaux passagers selon leur type, on pourrait calculer un taux de présence global pour ces passagers.

Les arbres de décision sont aussi utilisés pour déterminer les caractéristiques pouvant décrire la classe finale d'un objet caractérisé par un vecteur d'attributs. Ils sont composés de nœuds desquels partent des branches, à l'extrémité desquelles on retrouve pour chacune un nouveau nœud. Le nœud initial comporte l'ensemble de toutes les données de l'échantillon. On choisit ensuite un attribut en fonction duquel on fait la séparation de l'échantillon en deux (ou plusieurs) parties. Par exemple, si l'attribut sélectionné vaut 0 pour une donnée quelconque, on la place du côté gauche, si il vaut 1 ou plus, on la place du côté droit. La séparation doit diminuer l'hétérogénéité, ou l'entropie (Colton, S. 2004), du groupe initial au maximum, en fonction de la classe des données. On divise ainsi à nouveau chacun des deux groupes, jusqu'à ce qu'il n'y ait qu'une seule classe de données présente à l'intérieur de chaque groupe. L'arbre pour les passagers serait idéal lorsque dans chaque groupe on ne retrouverait que des passagers présents, ou encore, que des passagers absents, et donc tant qu'il y aurait les deux, on ferait une autre scission, à moins d'avoir épuisé la liste des attributs. Au plus bas niveau de l'arbre, les nœuds contenant les groupes homogénéisés deviennent des feuilles. Les caractéristiques de chaque groupe peuvent être lues en remontant tous les nœuds jusqu'au nœud initial, car chaque nœud contient la règle de séparation qui a été utilisée, c'est-à-dire l'attribut choisi et la valeur critique en fonction de laquelle on sépare. Les algorithmes d'arbres de décision biens connus sont C4.5, ID3, Public, pour ne nommer que ceux-ci.

Les règles d'association quant à elles ne peuvent pas classifier les passagers en tant que tels, mais elles servent plutôt à faire des remarques telles que lorsque l'attribut 1 est présent, cela implique que l'attribut 2 n'est pas présent pour 75% des observations, ou encore, lorsque l'attribut 3 vaut au moins 5, l'attribut 6 ne vaut pas plus que 2, dans 90% des cas. On aurait donc pu traiter la classe (présent ou absent) comme un attribut et tenter de voir quelles règles impliquant chaque classe sont récurrentes. Les méthodes pour détecter des règles d'association calculent de simples corrélations, mais elles offrent la possibilité de le faire avec plusieurs attributs à la fois ainsi que de choisir le niveau de confiance des règles et le nombre de termes maximum permis pour une règle.

Toutes ces méthodes ont fait leurs preuves et peuvent être utiles pour la classification des passagers. Toutefois, cette étude porte sur une nouvelle méthode n'ayant pas été utilisée à ce jour pour des problèmes comme celui de la survente. Il s'agit de la méthode LAD.

La méthode LAD émerge en 1988 avec les fonctions booléennes partiellement définies, développées par Peter L. Hammer et son équipe de l'Université de Rutgers. Initialement développée pour ne fonctionner qu'avec des attributs binaires, cette méthode a été élargie à n'importe quelles valeurs numériques au début des années 2000. La méthode LAD est très avancée aujourd'hui, et elle a été maintes fois prouvée efficace. Elle se distingue des autres par son caractère combinatoire et logique, et c'est l'approche retenue pour le projet.

1.2 Analyse logique de données en bref

Avant de détailler la méthode LAD, voici des concepts de base et des définitions qui en faciliteront la compréhension. D'abord, la base de données utilisée pour la méthode LAD est une série d'observations, chacune étant caractérisée par un vecteur d'attributs. Ces attributs proviennent des caractéristiques de bases. Une caractéristique est une

information que l'on peut obtenir au sujet du passager, tandis qu'un attribut est la forme utilisée pour représenter cette caractéristique lors de l'utilisation de la méthode LAD. On essaie d'avoir le moins de valeurs différentes possibles à l'intérieur d'un attribut (idéalement binaire), afin d'accélérer le temps de résolution, mais aussi parce que lorsque les attributs sont bien construits, les résultats de la méthode LAD sont nettement supérieurs. Par exemple, une des caractéristiques relevées pour décrire les observations de notre problème est la date de départ du vol. Pour cette caractéristique, l'attribut vaut 1 pour tous les lundis, mardis et mercredis, 2 pour les jeudis, 3 pour les vendredis, 4 pour les samedis et 5 pour les dimanches. Si le comportement des passagers est le même le lundi et le mardi, la valeur de l'attribut ne devrait pas être différente.

La méthode LAD est une technique de classification de données qui n'utilise pas les méthodes statistiques conventionnelles telles que celles décrites ci-dessus. Il s'agit d'une façon de faire du data mining intelligent ou dit supervisé. La base de données est d'abord scindée aléatoirement en deux parties, l'une pour l'apprentissage *(training*, TRA), l'autre pour la validation (*testing*, TES). L'algorithme cherche ensuite à construire des groupes de conditions, à partir des attributs, pour lesquels les observations correspondantes ont toutes le même comportement ou la même classe finale. Celle-ci peut être positive ou négative.

Un groupe de conditions qui couvre des données majoritairement positives (négatives) sera appelé un patron positif (négatif). Une fois que les patrons des deux classes identifiés, avec la deuxième partie de l'échantillon, on vérifie une à une les données pour établir tous les patrons auxquels elle répond. Il apparaît évident que si la donnée satisfait plusieurs patrons positifs (négatifs), mais aucun patron négatif (positif), elle sera classée automatiquement positive (négative). Si elle ne correspond à aucun patron, elle demeurera non classifiée. Par contre, si elle peut être couverte par des patrons provenant des deux types, on choisira de la classer dans le groupe qui est proportionnellement le

plus important, ou prépondérant selon tout autre mécanisme sélectionné (voir la section 3.4 pour plus de détails).

Les conditions ont la forme « attribut 1 < 5 », « attribut 2 > 2 », etc. Les patrons peuvent donc être composés d'un groupe de conditions sur différents attributs, comme dans l'exemple mentionné, ou encore, un intervalle compris entre deux valeurs d'un même attribut: « attribut 1 < 5 » et « attribut 1 > 2 ». Pour chaque attribut, les valeurs doivent être sous forme numérique. Il convient d'abord de convertir les valeurs des attributs catégoriques (qualitatifs) ou mixtes en nombre réels. Les points de coupures assurent ensuite la discrétisation de l'espace à l'aide d'un index cohérent: à l'intérieur de l'écart entre la plus petite et la plus grande valeur associées à chaque attribut, les valeurs sont renumérotées de 0 jusqu'à k. Les points de coupures servent aussi à diminuer le nombre de valeurs différentes à l'intérieur de chaque attribut. Les manières de les appliquer seront abordées plus en détails dans le chapitre 2. Il est à noter que pour les attributs binaires, le recours aux points de coupures est inutile.

Le nombre d'attributs ainsi que le nombre d'intervalles que cet attribut contient influencent le nombre de combinaisons possibles à explorer, augmentant implicitement le nombre de patrons qui peuvent être générés. En effet, lors de la construction d'un patron, l'algorithme doit choisir quel attribut utiliser, mais aussi à partir de quelle valeur les observations seront incluses ou exclues de ce patron; le nombre de possibilités augmente de façon exponentielle à chaque fois que l'on ajoute une valeur à l'intérieur d'un attribut. Il devient donc nécessaire de limiter le nombre de patrons à générer en fonction de quelques paramètres. Afin d'y parvenir, il suffit d'introduire des caractéristiques propres aux patrons, soient : l'homogénéité (on y fait parfois référence dans la littérature en termes de *risque*), la prévalence et le degré.

L'homogénéité se définit par le nombre d'observations positives incluses dans le patron, par rapport au nombre total d'observations satisfaisant ce même patron. Plus intuitivement, pour un patron positif (négatif), l'on recherchera de nombreuses données positives (négatives), et peu de données négatives (positives), donc une homogénéité élevée (faible). La prévalence quant à elle détermine si le patron couvre suffisamment de données pour être considéré comme significatif. Il s'agit du nombre d'observations positives (négatives) qui satisfont les conditions du patron, par rapport au nombre totales d'observations positives (négatives) dans l'ensemble de données servant à l'apprentissage. Finalement, le degré est le nombre de termes ou de conditions maximal à inclure dans le patron, i.e. le nombre de variables bornées. Plus le degré permis est élevé, plus il y a de combinaisons possibles, et ceci peut occasionner des difficultés au niveau du temps de résolution, ou du moins, des délais importants. Ce sont ces contraintes qui encadrent la formulation des théories: tous les patrons existants à l'intérieur de ces trois paramètres fixés sont énumérés. Le meilleur patron positif (négatif) possède une homogénéité élevée (faible), une prévalence élevée et son degré est petit.

Une fois que tous les patrons correspondant aux paramètres spécifiés ci-dessus sont générés (en anglais *pandect*), on peut en sélectionner un sous-groupe qui fera office de règles de classification. On résout généralement des problèmes de recouvrement afin de déterminer quels sont les meilleurs patrons à conserver. En effet, il est possible de réduire le nombre de patrons utilisés pour la classification, tout en maintenant la couverture de l'ensemble des données. Ceci n'affecte pas les résultats, car certains patrons peuvent être superflus ou redondants. C'est ce nouvel ensemble de patrons que l'on conserve pour le modèle, il s'agit de la théorie.

Afin de compléter le modèle, il ne reste plus qu'à créer le discriminant basé sur le sous-ensemble de patrons sélectionnés. Lorsqu'une nouvelle observation ne correspond à

aucun patron, elle demeure non classifiée. Si elle est couverte par des patrons positifs et/ou des patrons négatifs, il faut calculer le biais de l'observation à l'aide de l'équation discriminante; il s'agit d'une équation où chaque patron apparaît accompagné s'il y a lieu d'un poids à partir duquel on peut calculer un score pour chaque observation. Lorsque cette valeur est plus élevée que le seuil critique positif, l'observation est positive (passager présent). Lorsque le résultat est sous le seuil négatif fixé, elle est classifiée négative (passager absent).

Pour obtenir des patrons qui offriront les meilleurs résultats possibles, il est d'une importance capitale de bien discrétiser les attributs: il s'agit du choix des points de coupures. C'est de là que naissent les valeurs parmi lesquelles on peut choisir lors de la construction des patrons. Le choix de la combinaison des paramètres (homogénéité, prévalence et degrés) revêt également une grande importance, puisqu'il permet de contrôler le nombre et la qualité des patrons qui seront énumérés.

Il existe plusieurs outils potentiels avec lesquels il est possible d'appliquer la méthode LAD à une base de données: coder l'ensemble du processus, utiliser la trousse d'implémentation en C++ d'Eddy Mayoraz, ou encore, utiliser un logiciel maison mis sur pied par Sorin Alexe en 2002, au Rutcor Center, Rutgers University dont l'accès en mode démo est ouvert à tous. Nous avons opté pour ce logiciel afin d'accélérer le processus de développement.

Avant de passer à l'exploration des données et à la construction du modèle, il a été nécessaire de se familiariser avec cet outil afin de s'assurer qu'il était réellement possible de l'utiliser pour faire le travail requis et qu'il convenait aux besoins. Ce logiciel se compose de quatre fonctions principales : le prétraitement des données (téléchargement, séparation de l'échantillon), le positionnement des points de coupures, aussi appelé la discrétisation des données, la génération des patrons selon les paramètres fixés par

l'utilisateur, et finalement la construction du modèle ainsi que sa validation. Plus de détails seront donnés à la section 2.3 de ce mémoire.

1.3 Applications

Jusqu'à présent, la méthode LAD a été testée sur plusieurs bases de données existantes afin d'en démontrer la robustesse, mais cette approche a aussi été utilisée pour résoudre des nouveaux problèmes dans des domaines très diversifiés. Les premières bases de données sur lesquelles la méthode LAD a été appliquée sont tirées du *Repository of Machine Learning Data-bases and Domain Theories*, maintenues par l'Université de Californie, à Irvine. Le but était de démontrer la compétitivité de la méthode avec les méthodes plus traditionnelles.

Tableau 1.1: Comparaison de la méthode LAD avec d'autres méthodes

	Méthode LAD				**Meilleure approche trouvée**		
	50% Training		**80% Training**		**Moyenne**	**Écart type**	**% utilisé pour training**
Base de données	**Moyenne**	**Écart type**	**Moyenne**	**Écart type**			
Australian credit card	85,4	1,2	85,5	2,6	85,5	N/D	71%
Boston housing	84,0	1,6	85,2	3,0	83,2	3,1	80%
Breast cancer	96,9	0,9	97,2	1,3	96,2	0,3	80%
Congressional voting	96,2	1,1	96,6	1,8	95,6	N/D	66,6%
Diabetes	71,9	1,9	72,3	2,4	76	N/D	75%
Heart desease	82,3	1,7	83,8	5,2	80,6	3,1	N/D

Les moyennes sont calculées pour 20 essais, et la partie utilisée pour l'apprentissage est sélectionnée aléatoirement (le reste de la base de données servant à la validation) en

utilisant soit 50%, soit 80% de la base de données. La moyenne indique le nombre de fois que la méthode a bien classé une observation; donc 85 de moyenne signifie que 85% des données ont été classifiées avec exactitude, tandis que 15% ont été mal classées ou non classées.

Par exemple, le processus d'attribution d'une carte de crédit en Australie était basé sur quinze attributs: quatre binaires, cinq nominaux et six numériques. Le classement effectué par la méthode LAD (à 85,4% exact) se fait à l'aide d'un seul patron, contenant l'attribut 9 uniquement (degré 1). Fait intéressant: l'auteur soulève que l'utilisation de patrons additionnels dans le discriminant ne semble pas améliorer les résultats. Pour ce problème, les meilleurs résultats trouvés dans la littérature ont été obtenus par la méthode basée sur un arbre de décision proposée par Carte et Catlett (Boros, E., et al., 2000).

Les conclusions à tirer de ce tableau présentées plus en détails dans l'article de Boros, E. et al. (2000) sont que la méthode LAD est très concurrentielle sur plusieurs plans. Elle est robuste, stable, et facilement adaptable à des types différents de problèmes. Il est à noter que la partie de droite du tableau est composée de différentes méthodes (la méthode présentant les meilleurs résultats dans la littérature), et non pas d'une unique méthode universellement bonne pour chacune des applications étudiées. La méthode LAD quant à elle rivalise en termes d'exactitude avec chacune de ces méthodes bien connues en data mining.

Les premières études réalisées avec la méthode LAD ont démarré en parallèle et sont rapportées en profondeur dans l'article de Boros, E. et al. (2000) ainsi que dans celui de Hammer, A. B., Hammer, P. L. & Muchnik, I. (1999). Il s'agissait de la productivité des provinces en Chine, d'établir les endroits d'exploitation pour le pétrole et de tests psychométriques visant à identifier les gens à risque de dépression. Ces trois expériences sont sans contredits qualifiables de succès.

L'étude de la productivité de 29 provinces chinoises, pour une période de temps allant de 1985 à 1994 a été réalisée en fonction de huit attributs. Ces 290 observations ont été classifiées en fonction de leur productivité. Avec 12 patrons positifs et 15 patrons négatifs, on arrive à classifier les observations correctement à 92,7%. Certains patrons sont remarquables, l'exemple suivant est un patron positif qui est exact à 100% (sans aucune erreur de classification): la région n'est pas égale à 4 (l'ouest de la Chine contient des provinces plus défavorisées) ET le temps est plus grand ou égal à l'année 1988. C'est-à-dire que toutes les observations qui satisfont ce patron sont effectivement positives, et aussi, qu'aucune observation négative ne correspond à ce patron. Ceci met en évidence l'importance de la localisation géographique ainsi que le temps comme des facteurs déterminants de la productivité pour les entreprises en Chine.

Pour l'étude des sols afin de déterminer les zones d'exploitation du pétrole, deux bases de données ont été utilisées. Le premier ensemble comporte 1632 données, tandis que le deuxième en contient 2393. Pour le premier, il existe une théorie consistant en un seul patron de degré 3, dont le taux de prédiction s'avère exact à plus de 94%. Pour le deuxième, une théorie a été trouvée comportant cette fois-ci un seul patron de degré 2 et exact à 92,3%. Trois des sept mesures utilisées se sont avérées inutiles lors de cette étude; ce sont des attributs qui n'apparaissent dans aucun patron. Aussi, l'auteur fait remarquer qu'il existe un patron avec un seul attribut binaire - si la valeur de la caractéristique supérieure ou non à un seuil particulier - pouvant classifier correctement 90% des données. Ceci est d'autant plus surprenant que la meilleure corrélation décelée entre l'objectif et n'importe lequel de ces attributs n'est que de 0.63.

L'analyse des résultats des tests psychométriques de 280 patients a été réalisée à l'aide d'un test bien connu (Beck Depression Inventory, BDI) comportant 21 questions, à chacune desquelles le répondant peut mettre une valeur allant de 0 à 3. Les patients dont le résultat est supérieur à 18 sont qualifiés de dépressifs, tandis que les autres sont

normaux. La méthode LAD conclut que l'on peut utiliser des attributs binaires pour chacune des questions et que 16 questions suffisent pour le classement. La théorie retenue contient 12 patrons positifs et 11 patrons négatifs, lesquels comportent au maximum 3 degrés. Cette théorie est exacte à 90,2% et fait erreur dans 8,9% des cas, laissant moins de 1% de patients non classifiés. Non seulement cette classification est-elle jugée excellente par les experts en psychiatrie, mais elle permet aussi de distinguer plusieurs sous-groupes de dépressifs chez les patients selon quels patrons les recouvrent, pouvant nécessiter des traitements différents ou mieux adaptés. En effet, même si la méthode LAD est basée sur une classification binaire (positif/négatif), on peut utiliser plusieurs seuils critiques, permettant une stratification des observations en plus que deux groupes. Les patients ayant des scores très positifs peuvent ainsi être distingués de ceux qui ont des scores positifs, mais moins élevés, i.e. des cas moins lourds. Une autre utilisation des résultats est la construction de tests plus simples, comportant moins de questions, et pouvant être faits passer à un patient pour valider le score d'un autre test.

La méthode LAD a par la suite été employée pour des applications très variées. La méthode a été démontrée efficace à ce jour pour les suivantes: identification des risques de blocage de l'artère coronaire (Alexe, S., Blackstone, E. & Hammer, P. L., 2003), croissance des matériels biologiques polymériques (Abramson, S. D., Alexe, G., Hammer, P. L. & Kohn, J., 2005), diagnostiques de problèmes pulmonaires (Boros, E. et al., 2000), repérage de lymphomes diffus à grandes cellules B (Alexe, G., Alexe, S., Axelrod, D., Hammer, P. L. & Weissman, D., 2005), contrôle des pièces défectueuses sur les appareils aériens (maintenance) (Bennane, A. & Yacout, S., 2009). Cette dernière application a été publiée tout récemment, en novembre 2009 et a été développée à l'École Polytechnique.

Compte tenu des succès de cette approche dans divers contextes, il serait intéressant de voir comment se comporte cette approche pour la détection des passagers qui seront

présents ou non à la porte d'embarquement. La méthode LAD n'a pas encore été employé pour la prévision de ce genre de phénomène aléatoire du comportement humain. À l'heure actuelle, les modèles de prévisions des ventes en transport aérien sont ajustés selon les taux de présence historiques afin de vendre quelques sièges supplémentaires.

1.4 La survente en transport aérien

Les pratiques de survente de sièges varient largement d'une compagnie aérienne à l'autre. Dans la plupart des compagnies aériennes, on utilise une formule probabiliste qui tient compte de la pénalité encourue par un refus d'embarquement à un client confirmé, appelé en anglais « denied boarding », de la capacité de l'appareil, ainsi que de la probabilité qu'un passager se présente (Hillier F. S. et al., 1998). Cette probabilité est le taux de présence, généralement calculée à partir de l'historique du vol. Une autre formule encore plus simple est fréquemment utilisée :

$$Capacité\ \ autorisée = \frac{Capacité\ réelle}{1 - taux\ d'absence}$$

Peu importe le calcul choisi, le taux de présence (ou d'absence) est un input majeur du problème de gestion d'inventaire aérien. De plus, le phénomène de la survente est utilisé dans d'autres secteurs où les pertes de revenus supplémentaires potentiels existent lorsque des clients sont absents, notamment en transport ferroviaire et dans l'industrie hôtelière.

Bien entendu, on doit aussi tenir compte du facteur de risque que l'entreprise est prête à encourir. Par exemple, la compagnie WestJet a choisi de ne faire aucune survente. Il s'agit d'une décision stratégique sur le plan marketing, car la compagnie fait ainsi sa publicité autour de cette politique en garantissant à ses clients qu'ils ne se feront jamais refuser l'embarquement.

À l'opposé, chez Air Canada, on cherche de manière continue à développer des moyens pour améliorer les prédictions d'absences, afin d'ajuster la survente en conséquence. L'article de Lawrence, R. D., Hong, S. J. & Cherrier, J. (2003) le démontre: une méthode d'ajustement des prédictions basée sur les régressions linéaires des attributs de passagers a été prouvée efficace. L'idée d'analyser les passagers en fonction de leurs attributs n'est donc pas nouvelle. Cet article et un des auteurs employé d'Air Canada ont d'ailleurs été consultés pour établir la liste d'attributs potentiellement influents. Un autre groupe d'auteurs, Gorin, T., Brunger, W. G. & White, M. (2006) ont également évalué la valeur d'ajuster le modèle de prédiction en fonction des trois attributs les plus influents selon leur article, soient la proportion des passagers sur le vol qui sont sur leur segment de retour, la proportion de passagers ayant des billets électroniques et la répartition des passagers entre ceux qui sont locaux, par rapport à ceux qui arrivent d'un autre vol (connexion).

Finalement, sur le site du *Airline Group of the International Federation of Operational Research Societies* (AGIFORS), on peut voir en un coup d'œil qu'il existe relativement peu de documentation ou de présentations ayant eu lieu dans des congrès au sujet de la survente ou des prévisions de passagers absents. Malgré que la survente de sièges soit une source non négligeable de revenus équivalent à des millions de dollars, cette question n'est pas abordée très en détails.

CHAPITRE 2 ANALYSE LOGIQUE DE DONNÉES ET SURVENTE

Dans ce chapitre, le parallèle entre la méthode LAD et le problème des passagers absents ainsi que la manière d'appliquer cette méthode sont d'abord explicités. Les sous-sections suivantes décrivent respectivement les caractéristiques potentiellement utiles pour déceler les risques de présence ou d'absence d'un passager, le travail de préparation nécessaire pour rendre les données compatibles avec la méthode LAD, et finalement, les bases du logiciel générateur de patrons, le LAD Datascope V2.0 développé par Alexe Sorin. Avant d'aborder le cœur du sujet, voici quelques spécifications:

- un passager présent est une observation positive;
- un passager absent est une observation négative;
- les attributs sont élaborés à partir d'une ou de plusieurs caractéristiques.

2.1 Caractéristiques des passagers

Il faut déceler quelles caractéristiques des voyageurs peuvent influencer leur tendance à se présenter ou à s'absenter. Il y a certains facteurs qui semblent intuitivement pertinents, tels que la classe du billet et le nombre de personnes qui voyagent ensemble. Effectivement, certains billets sont remboursables en cas d'absence, tandis que d'autres ne le sont pas. On pourrait donc s'attendre à observer un plus grand taux d'absence pour ce premier type de billet. Aussi, on pourrait s'attendre à ce qu'une personne seule soit plus prédisposée à s'absenter qu'un des membres d'une famille voyageant à quatre.

L'élaboration de cette première liste devait être la plus vaste possible, en sachant tout de même qu'on ne pourra pas utiliser ou extraire toutes ces informations pour diverses raisons (disponibilité de l'information, confidentialité). La liste a donc été conçue à l'aide de plusieurs acteurs ayant œuvré dans ce domaine chez Air Canada et elle est basée

sur leur expérience antérieure par rapport aux facteurs qui peuvent influencer le comportement d'un passager. Les différentes bases de données ont aussi été parcourues afin de relever un maximum d'idées de caractéristiques qui pourraient avoir un impact quelconque. Il existe deux types de caractéristiques : celles qui sont relatives aux vols, qui sont donc identiques pour tous les passagers sur un même vol, et celles qui sont relatives aux individus, plus précisément à leur réservation. Ces caractéristiques deviendront des attributs lors de la phase de préparation des données.

Caractéristiques relatives aux vols

Durée du vol
Origine du vol
Destination du vol
Heure de départ (locale)
Heure d'arrivée (locale)
Date de départ
Date d'arrivée
Journée de la semaine
Alternatives possibles de vols ou de routes
Fréquence des vols sur le segment étudié

Ce genre de caractéristiques peut influencer le comportement du passager de manière générale. On peut penser par exemple qu'un vol ayant lieu très tôt le matin, et pour lequel il existe un vol vers la même destination 30 minutes plus tard comportera plus d'absents. Il n'est pas exclu qu'une caractéristique par elle-même n'apporte pas beaucoup d'information, mais lorsque jumelée à une autre, devienne un bon outil de prévision. La richesse des patrons à venir réside dans la combinaison de plusieurs de ces facteurs.

Caractéristiques relatives aux passagers

Genre

Âge

Nombre de personnes voyageant ensemble

Motif du voyage (affaires ou tourisme)

Membre d'un programme de fidélisation (Aéroplan)

Nombre de changements effectués depuis la première journée de la réservation (ajout/retrait de personnes, changement de vol, etc.)

Date du dernier changement effectué

Historique des absences de ce passager

Classe de réservation

Nombre de jours entre la date d'achat du billet et la date du départ (avance)

Origine de l'itinéraire du passager

Destination de l'itinéraire du passager

Durée totale de l'itinéraire

Présence d'un numéro de billet

Présence d'un numéro de billet électronique

Billet aller-retour

Passager sur la partie retour de son itinéraire

Durée du séjour

Passager arrivant d'un autre vol (connexion)

Temps de connexion entre le vol précédant et celui étudié

Comportement habituel du vol précédant le transfert (historique de retard)

Présélection du siège

Paiement à l'avance pour bagages excédentaires

Options prépayées (accès au « Mapple Leaf lounge », choix de repas, etc.)

Embarquement

Bien que les liens entre la plupart de ces caractéristiques et le comportement du passager semblent plutôt évidents, certains le sont moins. Par exemple, le nombre de changements effectués peuvent indiquer un passager plutôt incertain, ou au contraire, plus prévoyant. De plus, certaines de ces informations sont difficiles à retrouver, ou encore, confidentielles, comme l'historique des absences du passager ou son âge.

Finalement, il est à noter qu'il s'agit d'un ensemble de départ seulement et que tous ces attributs peuvent potentiellement influencer le champ qui nous permet de savoir si le passager était présent à la porte d'embarquement; l'objectif consiste alors à déterminer quels attributs sont les plus pertinents. La caractéristique « Embarquement » n'en est pas une en tant que telle, il s'agit plutôt de l'indicateur qui nous permet de savoir si le passager est de type positif (présent) ou négatif (absent).

2.2 Préparation des données

En plus d'extraire toutes ces informations, il faut adapter les données disponibles chez Air Canada afin de pouvoir faire fonctionner le logiciel et d'obtenir une solution qui soit utilisable. Dans ce chapitre, le processus non négligeable de collecte de données, ainsi que l'adaptation nécessaire pour les valeurs et la forme de ces données sont décrits.

Plusieurs bases de données de très grande taille existent chez Air Canada. Elles proviennent de systèmes qui fonctionnent en parallèle : le système des réservations, le système de récupération des données post-départ relié à l'aéroport, le système des informations au sujet des vols, pour ne nommer que ceux-ci. Heureusement, récemment, une application dans Business Object (BO) a été construite à l'interne afin de jumeler les différents systèmes. Cette application permet de faire systématiquement les liens entre les tables d'une base de données à l'autre, afin d'obtenir des données cohérentes. Tous les champs de la plupart des systèmes existants sont listés sur BO. Il suffit de sélectionner le

ou les champs désirés. Suite à cette sélection, on peut visualiser le code source de la requête en langage SQL et puis finalement lancer la requête à partir d'un générateur de données appelé TOAD.

Dans BO, le nombre de champs est très élevé. Certains d'entre eux sont redondants ou incomplets et pour d'autres, il est difficile d'en retracer la signification. À partir de la liste proposée du départ, nous avons tenté d'amasser un maximum d'attributs. Il est à noter que certains de ces attributs n'existaient pas à « l'état pur » et qu'il fallait parfois mettre en relation plusieurs champs afin de déduire la valeur de cet attribut. D'autre part, certains attributs ont été abandonnés par souci de simplicité.

Après avoir déterminé les champs retenus, il faut faire le prétraitement des données recueillies (simplification, numérisation et discrétisation) pour profiter de la meilleure utilisation du logiciel possible. L'objectif consiste à obtenir la meilleure séparation possible des données positives et des données négatives, mais en utilisant le moins de séparations possibles et en conservant un nombre suffisant de données à l'intérieur de chacune. Effectivement, on ne veut pas séparer entre chaque donnée, mais à l'opposé s'il n'y a pas suffisamment de points qui départagent les données positives des négatives, les patrons n'auront pas de bonnes homogénéités. Les données ont été extraites pour une période se limitant à un mois, plus précisément le mois de mars 2009, pour tous les vols allant de Vancouver vers Calgary, commercialisés par Air Canada.

La liste finale des attributs retenus ainsi que chacune des valeurs qui peuvent être associées à ces attributs est présentée ci-dessous. Elle a été modifiée plusieurs fois au cours du projet, soit par l'élimination de certains attributs, soit par la diminution du nombre de valeurs différentes à l'intérieur d'un attribut. La plus grande partie du nettoyage a été fait à l'aide d'Excel et d'Access. Plus de détails sont présentés sur l'épuration des attributs à la section 3.2.

Tableau 2.1: Attributs retenus, valeurs et descriptions

Attributs	**Valeurs**	**Descriptions**
Présence de billet	0	Aucun
	1	Présence d'un numéro de billet dans le PNR*
Jour de la semaine	1	Dimanche
	2	Lundi, mardi ou mercredi
	3	Jeudi
	4	Vendredi
	5	Samedi
Numéro de segment	1	Premier segment de vol de l'itinéraire total
	2	Deuxième segment de vol
	3	Troisième, quatrième ou cinquième segment de vol
	4	Sixième segment de vol et plus
Avance	0	Réservation effectuée 60 jours ou moins avant le départ
	1	Réservation effectuée 61 jours ou plus
Options	0	Aucune choisie par le passager
	1	Choix d'au moins une des sept «go-options»
Genre	0	Aucun noté dans le PNR (incomplet)
	1	Homme ou femme
Voyageur fréquent	0	Non
	1	Accumulation de points Aéroplan
Classe de réservation	1	Tango
	2	Tango plus
	3	Latitude
	4	Affaires
	5	Points Aéroplan
Origine de l'itinéraire total	1 à 10	Selon la région géographique
Destination de l'itinéraire total	1 à 10	Selon la région géographique
Point de vente du billet	1 à 10	Selon la région géographique
Billet électronique	0	Non
	1	Présence d'un numéro de billet électronique dans le PNR
Billet acheté aller-retour	0	Non
	1	Si l'origine est égale à la destination dans l'itinéraire total

OD_type	4	Vols non inclus dans AC Inc.
	6	Vols inclus dans AC Inc. mais pas dans AC seulement
	7	Vols inclus AC seulement (et donc AC Inc. aussi)
Heure de départ	1	De 6h00 à 7h59
	2	De 8h00 à 9h59
	3	De 10h00 à 12h59
	4	De 13h00 à 15h59
	5	De 16h00 à 17h59
	6	De 18h00 à 19h59
	7	De 20h00 à 21h59
	8	De 22h00 à 5h59
Nombre de personnes	1	1 seule personne enregistrée dans le PNR
	2	2 personnes enregistrées
	3	3 personnes et plus
Embarquement	0	Non
	1	Oui

* Le PNR est le « Passenger Name Record », il s'agit du dossier qui correspond à la réservation.

2.3 Utilisation du logiciel

La façon dont le logiciel fonctionne est décrite dans cette section. Pour chacun des onglets principaux, ce qui peut être fait ainsi que les différents paramètres sur lesquels on peut jouer pour tenter d'améliorer les résultats sont présentés. Il existe aussi des commandes que nous ne mentionnons pas par souci de simplicité. La figure qui suit montre la fenêtre du menu principal qui apparaît lors de l'ouverture.

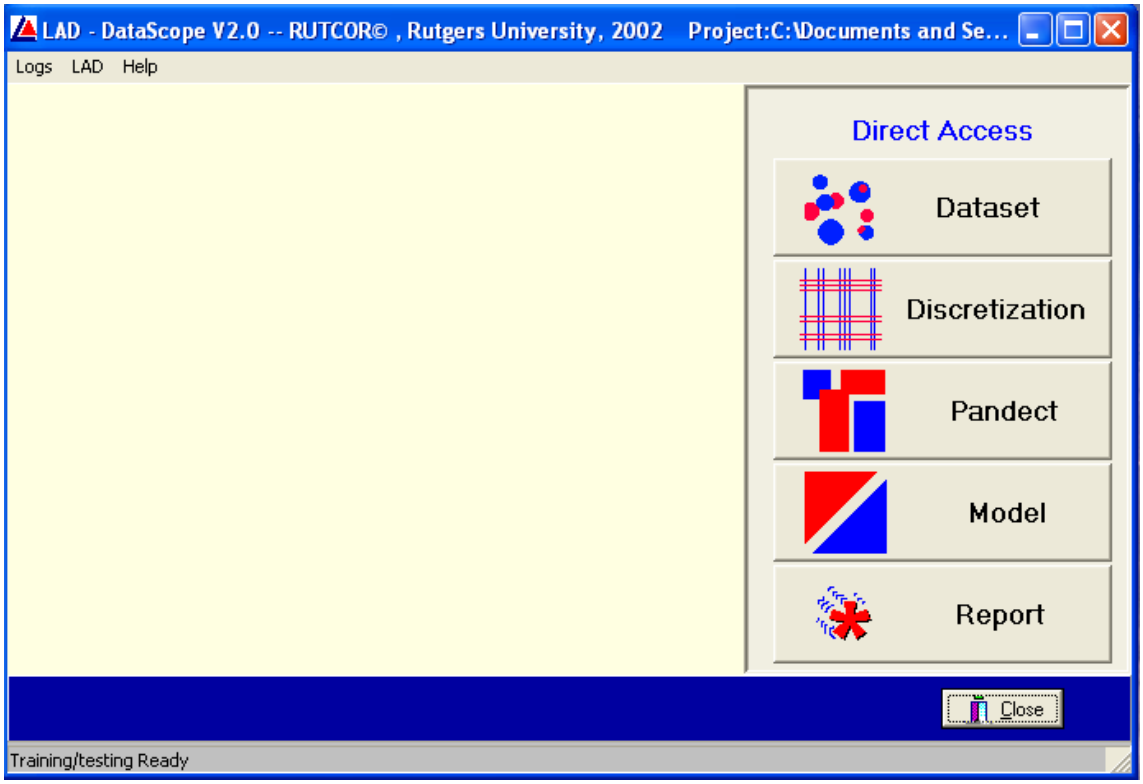

Figure 2.1: Interface avec les principaux onglets du LAD Datascope V2.0 RUTCOR ©

La première étape consiste à télécharger les données dans le logiciel. La meilleure façon de faire consiste à ouvrir les données en fichier texte directement à partir du bouton «Load data» dans l'onglet « Dataset ». Il est ensuite possible de répartir les données en deux ensembles, le premier qui servira pour le « training », c'est-à-dire la lecture des données pour la création de patrons, tandis que la deuxième partie servira au « testing », ou à la validation. Plusieurs manières de diviser les données sont proposées: aléatoirement (« random »), par « cluster » ou « leave one out » qui consiste à ne laisser qu'une seule donnée pour le test de validation et mettre toutes les autres en « training ». Avant d'utiliser une de ces méthodes, il convient d'appuyer sur le bouton « Mix up » qui, comme son nom l'indique, mélange toutes les données aléatoirement. Le mode de séparation retenu pour le projet est aléatoire. Avec cette méthode, l'utilisateur détermine

les proportions respectives des données dans l'ensemble d'apprentissage et de validation. Dans le cadre de ce projet nous avons opté pour une répartition 50%-50%.

Une fois les données téléchargées dans le logiciel et séparées en deux, le deuxième onglet principal du menu permet de passer à la phase suivante de la méthode LAD: la discrétisation des données. Cette phase est cruciale, car tous les patrons qui seront générés le seront à partir des valeurs qui limitent chacun des intervalles créés. Le but est donc de séparer là où il y a des changements significatifs de comportement des passagers. Encore une fois, le LAD Datascope propose plusieurs manières de procéder: en « bins », où les données seront séparées en nombre égal dans chacun des ensembles, « cluster-based », où les données seront séparées en fonction des principes de segmentation présentés dans le chapitre 1, en intervalles, où il y aura la même distance entre chaque valeur, ou par séparation, une méthode qui a pour but de séparer le plus possible les données négatives des données positives. Sauf pour la méthode de séparation, l'utilisateur doit indiquer de nombre de points de coupures, c'est-à-dire le nombre de séparations souhaité, pour chaque attribut. Pour la méthode par séparation, ce dernier n'a qu'à indiquer les nombres minimum et maximum de points de coupures qu'il désire utiliser.

Finalement, nous avons effectué la sélection des points de coupures en utilisant Excel, et renommé toutes les valeurs en fonction des indexes ainsi créés avant d'utiliser le logiciel. La discrétisation dans le logiciel s'est donc limitée pour nous à séparer chaque attribut en autant d'intervalles de longueur 1, qu'il y avait de valeurs différentes.

Le LAD Datascope propose également plusieurs outils, dont deux très utiles. Le premier, « Grid report », permet à l'usager de voir et même de modifier les points de coupures établis de manière plus systématique à l'étape précédente. Le deuxième, « Check grid quality », donne l'entropie de la grille, c'est-à-dire un indicateur de la qualité de la

séparation. Plus l'entropie est faible, meilleure est la séparation entre les données positives et les données négatives.

Lorsque la phase de discrétisation est terminée, c'est le moment de construire des patrons. Le terme « pandect » désigne l'ensemble des patrons générés, mais il ne s'agit pas nécessairement de tous ceux qui sont retenus, qu'on appelle plutôt la théorie ou le modèle. Il est important de trouver le plus grand nombre possible de « bons » patrons lors de cette étape, car par la suite, ce sont parmi ces patrons que seront sélectionnés ceux qui feront office de modèle. C'est durant cette phase qu'intervienne les notions d'homogénéité, de prévalence et de degrés présentées plus haut. En plus de fixer ces trois paramètres, l'utilisateur a la possibilité d'imposer une limite sur le nombre maximal de patrons à générer. Ceci sert à avorter le processus s'il s'étend trop, généralement causé par un mauvais choix de paramètres (trop vastes). Le LAD Datascope propose deux manières différentes de générer les patrons avec les mêmes paramètres: en cônes et en intervalles.

Compte tenu que nous n'avons pu obtenir une description convenable de la méthode avec les intervalles et que la méthode des cônes est systématique, cette dernière a été retenue. Son fonctionnement est très simple. Le logiciel essaie de faire des combinaisons utilisant au maximum n termes, où n désigne le nombre de degrés choisis par l'utilisateur. Il considère d'abord le premier attribut et sa première valeur. Si cette combinaison satisfait les paramètres d'homogénéité et de prévalence elle forme alors un patron et on passe alors à la deuxième valeur de l'attribut 1. Si toutefois cette combinaison ne peut être considérée comme un patron, le LAD Datascope ajoute un autre terme (la première valeur du deuxième attribut) à la combinaison. Si par exemple le nombre de données inclues est déjà trop petit pour que ce patron puisse satisfaire la prévalence souhaitée, le logiciel passera alors immédiatement à la combinaison suivante. Le logiciel balaie ainsi toutes les combinaisons afin de déterminer lesquelles peuvent former des patrons.

Attributs utilisés pour l'exemple:

1. Présence de billet

    0: vide

    1: présence d'un numéro de billet

    Point de coupure = 1, en réalité le « 1 » signifie plutôt « le premier point de coupure », et il se trouve à 0,5

2. Jour de la semaine

    1: dimanche

    2: lundi, mardi et mercredi

    3: jeudi

    4:vendredi

    5: samedi

    Points de coupures = 1 (1,5), 2 (2,5), 3 (3,5) et 4 (4,5)

Exemple d'ordre des combinaisons visitées par le logiciel, pour 2 degrés:

Présence de billet ≤ 1

Présence de billet ≤ 1, Jour de la semaine ≤ 1

Présence de billet ≤ 1, Jour de la semaine > 1

Présence de billet ≤ 1, Jour de la semaine ≤ 2

Présence de billet ≤ 1, Jour de la semaine > 2

Présence de billet ≤ 1, Jour de la semaine ≤ 3

Présence de billet ≤ 1, Jour de la semaine > 3

Présence de billet ≤ 1, Jour de la semaine ≤ 4

Présence de billet ≤ 1, Jour de la semaine > 4

Présence de billet > 1

Présence de billet > 1, Jour de la semaine ≤ 1

Présence de billet > 1, Jour de la semaine > 1

Présence de billet > 1, Jour de la semaine ≤ 2

Présence de billet > 1, Jour de la semaine > 2

Présence de billet > 1, Jour de la semaine ≤ 3

Présence de billet > 1, Jour de la semaine > 3

Présence de billet > 1, Jour de la semaine ≤ 4

Présence de billet > 1, Jour de la semaine > 4

Pour cet exemple, si nous avions permis au logiciel d'aller jusqu'à 3 degrés, le troisième terme du patron aurait pu être un point de coupure différent du deuxième terme, mais tiré du même attribut, par exemple:

Présence de billet ≤ 1, Jour de la semaine > 1, Jour de la semaine ≤ 4

Donc en réalité, ce patron inclue tous les passagers n'ayant pas de numéro de billet émis, voyageant un des jours suivant : lundi, mardi, mercredi, jeudi et vendredi.

Une fois que tous les patrons encadrés par les paramètres fixés sont générés, il s'agit de mettre sur pied le modèle en tant que tel. Encore une fois, le LAD Datascope offre plusieurs choix : on peut reprendre tous les patrons générés lors de l'étape précédente, prendre les 250 plus importants seulement, ou encore, en sélectionner un sous-ensemble en résolvant un problème de recouvrement (avec un algorithme glouton, par exemple). C'est aussi lors de cette étape qu'il faut fixer les poids pour chacun des patrons. Le logiciel permet d'imposer des pondérations égales pour tous, ou de choisir la prévalence, mais en pratique il existe d'autres manières de faire cette étape. Normalement, c'est aussi

lors de cette phase qu'il faut fixer les seuils critiques pour les scores des observations à partir desquels une observation sera finalement classée positive ou négative.

Pour notre projet, nous avons utilisé la valeur critique 0 pour les deux types de données (positives et négatives), et des pondérations équivalentes aux prévalences relatives. Cela signifie que s'il y a 15 patrons positifs et qu'une donnée correspond à 5 de ces patrons, elle obtient le score +0,33. S'il y a également 2 patrons négatifs et que cette même donnée satisfait un des deux, elle obtient un score de -0,5. On additionne les deux scores pour obtenir le total. Puisqu'il est inférieur à 0, cette donnée est classée négative en définitive.

Parmi les informations présentées dans les rapports que génère automatiquement le LAD Datascope, on trouve entre autres le nombre d'observations positives et négatives recouvertes par chaque patron, la fréquence d'apparition de chacun des attributs dans le modèle, le score calculé pour chaque donnée ainsi que son classement réel et celui prédit. À l'annexe 1, le lecteur trouvera le guide de l'utilisateur produit par Alexe Sorin (disponible en anglais seulement).

CHAPITRE 3 EXPLORATION

La phase exploratoire du projet consiste à préparer les données pour qu'elles puissent être traitées de façon efficace par le logiciel et d'appliquer la méthode LAD pour que l'on puisse soutirer la meilleure information possible. Cette phase est non négligeable dans ce projet compte tenu des limites imposées par le logiciel, du peu de documentation existant à son sujet et finalement des mécanismes à mettre en place afin de traiter d'aussi grandes banques de données. La phase exploratoire s'est donc étalée du début jusqu'à la toute fin du projet, les questions suscitées évoluant en même temps que les différents essais.

Ce chapitre se divise en quatre sections: la première contient des conclusions générales qui sont apparues au fil des essais, la deuxième explique comment nous sommes parvenus à la liste d'attributs finale présentée dans le chapitre 2, la troisième résume les essais ayant portés sur le nombre de degrés à utiliser et finalement la dernière section présente des essais pour lesquels les résultats obtenus nous ont semblé moins concluants.

3.1 Remarques générales

Le nombre d'observations comprises dans les banques de données d'Air Canada est énorme et celles-ci ne peuvent pas toutes être prises en compte dans le LAD Datascope puisque ce dernier limite à 16000 le nombre total d'observations. Cette limite nous a forcés à ajuster le processus d'échantillonnage. Pour notre expérimentation, nous avons donc sélectionné les observations relatives à un seul marché, pour tous les départs d'un mois donné. Cette restriction permet de limiter la taille des populations en-deçà de 40 000 observations, ainsi, les 16 000 observations constituent un échantillon représentatif de la population étudiée.

Pour nos tests, la population étudiée est composée de l'ensemble des passagers voyageant sur le segment Vancouver-Calgary durant le mois de mars 2009. Cette population représente 38 501 passagers; l'échantillon comprend environ une donnée sur trois. De ces données initiales, nous avons éliminé tous les passagers de la catégorie « Transfer to ». Cette catégorie désigne des passagers qui étaient préalablement enregistrés sur un autre vol, et qui, une fois sur place à l'aéroport, pour une raison ou pour une autre (préférence du passager, retard, type de billet, etc.), sont transférés sur le vol à l'étude. On ne veut pas tenir compte du profil de ces passagers qui sont assurés d'être présents et qui représentent des cas exceptionnels. Par contre, sur leur vol initial, ils apparaîtront comme des passagers absents.

Une fois que les essais se sont avérés plus stables et significatifs pour le premier échantillon, nous avons sélectionné des échantillons regroupant 16 000 observations différentes parmi la même population des 38 501 passagers voyageant en mars 2009, afin de vérifier la reproductibilité des résultats et la stabilité de l'approche.

Ceci modifiait quelque peu les paramètres à fixer pour obtenir les premiers patrons (paramètres très strictes), mais cela ne changeait que très peu les patrons construits en plus grande quantité (paramètres plus permissifs). Les résultats sont donc reproductibles en faisant varier l'échantillon à l'intérieur de la population seulement si on utilise des patrons en quantité suffisante, par contre lorsque que les patrons créés satisfont tout juste la valeur limite imposée pour leur création, des différences assez substantielles peuvent être observées.

Il en va de même quant à la répartition des données entre l'apprentissage (TRA) et la validation (TES). Elle est systématiquement différente lors de chaque ouverture du logiciel, donc presque chaque journée (même en gardant les mêmes proportions). Les résultats en étaient parfois affectés, surtout lorsque très peu de patrons sont créés et que

ceux-ci satisfont tout juste aux valeurs minimales exigées pour la création des patrons. Il peut arriver qu'une ou deux observations fassent la différence entre la création ou non d'un patron lorsque les paramètres utilisés sont très strictes.

Pour illustrer ceci, prenons l'exemple d'une prévalence fixée à 5% et d'une homogénéité fixée à 100% pour le type positif. Si la partie TRA contient 1000 données, cela revient à dire que les patrons doivent être vrais pour au moins 50 observations positives et strictement aucune observation négative. Imaginons qu'un patron soit vrai pour exactement 50 données. Si la répartition était faite différemment à une seule donnée près, cela pourrait porter à 49 le nombre d'observations positives, et on perdrait ce patron car il ne satisferait plus la prévalence minimale. D'autre part, il pourrait y avoir une donnée négative supplémentaire qui exceptionnellement réussisse à satisfaire le patron, donc l'homogénéité de 100% serait perdue; ce patron ne serait pas créé non plus. Finalement, nous avons trouvé une manière d'enregistrer la séparation de l'échantillon afin d'assurer une certaine cohérence entre les différents essais effectués d'une journée à l'autre. Encore une fois, la méthode LAD est reproductible seulement si l'on ne se situe pas aux limites extrêmes des paramètres, peu importe la répartition retenue.

Cette répartition est en fait une séparation aléatoire de l'échantillon à 50%-50%. Les séparations à 90%-10% ou 10%-90% ont également été testées, mais elles n'offraient pas des résultats aussi intéressants et présentaient plus d'instabilité. Effectivement, lorsque nous mettions trop (ou pas suffisamment) de données dans la partie qui sert à l'apprentissage, cela avait un impact. Si on exige par exemple une prévalence de 5% et qu'il n'y a que 10% des données (1500) dans la partie apprentissage de la base, cela nécessite 75 données d'un même type, pour former un patron. Si, de plus, l'homogénéité positive est fixée à 99%, en réalité, on veut 75 données positives sur un total de 75,75 données totales pour avoir un patron valide. Par contre, si le patron contient 76 données totales, 75 données positives équivaut à une homogénéité de seulement 98,6% donc le

patron serait rejeté. Il faudrait que les 76 données soient positives. Pour le logiciel, avec cet exemple, que l'homogénéité soit de 99% aura le même effet que si elle était de 100%. À l'inverse, s'il y a 90% des données (13 500), avec une prévalence de 10%, on voudrait 1 350 données pour construire un patron. Il y a une grande différence entre une homogénéité de 100% et 99%: 13 données (1363,6-1350 = 13,6).

Plusieurs centaines d'essais ont été faits en faisant varier l'homogénéité, la prévalence et le nombre degrés afin d'obtenir les meilleurs résultats. Voici quelques résultats qui ont pu être tirés parmi les essais réalisés avec un degré égal à 4 (la section 3.3 porte sur le nombre de degrés), et pour la classe égale à 5, i.e. les passagers voyageant avec leurs points Aéroplan accumulés (voir section 3.2) afin de mieux montrer comment les paramètres influencent les résultats.

Exemple 1:

Homogénéité positive: 100% - négative: 10%
Prévalence: 11%
Nombre de patrons générés: 2 positifs, 0 négatif

Tableau 3.1: Résultats de l'exemple 1 en valeurs absolues

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	398	0	2061
Négatives	2	0	266

Ce tableau indique que 398 passagers ont été classifiés présents correctement, mais que 2 passagers absents ont été classifiés comme présents aussi. Il n'y a aucune observation classée négative, car il n'y avait pas de patrons de ce type. Le reste des données demeurent non-classifiées.

Tableau 3.2: Résultats de l'exemple 1 en pourcentages des données du même type

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	16,19%	0%	83,81%
Négatives	0,75%	0%	99,25%

Le deuxième tableau donne la répartition des données de chaque type en pourcentage entre les trois groupes. Pour les positifs par exemple, 16,19% des observations de ce même type (ou 398 sur 2061+398) sont classifiées correctement et le reste est non classé.

Tableau 3.3: Résultats de l'exemple 1, par groupes générés

	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	99,50%	---	88,57%
Prévalence	14,67%	0%	84,33%

Pour l'exemple 1, avec ces 2 patrons, on réussit donc à générer un groupe comprenant 14,67% du total des observations (ou 400 sur 2327+400), et pour lesquelles on observe un taux de présence de 99,5% (ou 398 sur 400). Le groupe des non classés contient donc toutes les observations restante, soit 84,33% des observations totales, puisque le groupe des négatifs est vide. Le taux de présence de ce dernier groupe chute à 88,57% (ou 2061 sur 2327). L'exemple 1 est repris ci-dessous, en diminuant la prévalence de 11% à 10%.

Exemple 2:

Homogénéité positive: 100% - négative: 10%
Prévalence: 10%
Nombre de patrons générés: 13 positifs, 0 négatif

Tableau 3.4: Résultats de l'exemple 2 en valeurs absolues

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	607	0	1852
Négatives	10	0	258

Tableau 3.5: Résultats de l'exemple 2 en pourcentages des données du même type

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	24,68%	0%	75,32%
Négatives	3,73%	0%	96,27%

Tableau 3.6: Résultats de l'exemple 2, par groupes générés

	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	98,38%	---	87,77%
Prévalence	22,63%	0%	77,37%

Puisque la prévalence minimum requise est plus petite dans ce deuxième exemple, cela est l'équivalent de générer des patrons pour lesquels moins de données doivent satisfaire les conditions, donc qui sont peut-être moins significatifs ou importants. Toutefois, puisque les paramètres sont plus permissifs, il y aura plus de patrons. On peut donc observer que le taux de présence à l'intérieur du premier groupe est légèrement diminué, mais qu'on englobe cette fois-ci beaucoup plus de données dans ce groupe, soit 22,63% des passagers.

Avec ces mêmes paramètres, il faut diminuer la prévalence jusqu'à 7% afin que les premiers patrons de type négatif apparaissent. Évidemment, le nombre de patrons positifs continue d'augmenter à chaque point de prévalence que l'on diminue.

Exemple 3:

Homogénéité positive: 100% - négative: 10%

Prévalence: 7%

Nombre de patrons générés: 54 positifs, 8 négatifs

Tableau 3.7: Résultats de l'exemple 3 en valeurs absolues

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	1098	6	1355
Négatives	27	16	225

Tableau 3.8: Résultats de l'exemple 3 en pourcentages des données du même type

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	44,65%	0,24%	55,10%
Négatives	10,07%	5,97%	83,96%

Tableau 3.9: Résultats de l'exemple 3, par groupes générés

	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	97,60%	27,27%	85,76%
Prévalence	41,25%	0,81%	57,94%

Malgré une légère diminution du taux de présence dans le groupe de type positif, on réussit tout de même à rassembler 41,25% des données. Pour le groupe de type négatif, le taux de présence est très bas, soit 27,27%. Malheureusement, ce groupe est peu utile puisque qu'il ne contient que 0,81% des passagers. Il faut toutefois noter que le deuxième groupe (type négatif) ne rassemblera jamais un très fort pourcentage des passagers puisque les absents forment un pourcentage limité de la population (entre 5% et 25%, selon le produit). C'est donc le plafond de la prévalence pour ce groupe.

L'exemple 4 reprend la même prévalence que l'exemple 1 (11%), mais cette fois-ci, nous permettons aux patrons de ne pas être 100% purs. Les parons positifs pourront donc englober des données négatives (au maximum 1%) et les patrons négatifs pourrons inclure jusqu'à 15% de données positives.

Exemple 4:

Homogénéité positive: 99% - négative: 15%
Prévalence: 11%
Nombre de patrons générés: 21 positifs, 0 négatifs

Tableau 3.10: Résultats de l'exemple 4 en valeurs absolues

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	1107	0	1352
Négatives	32	0	236

Tableau 3.11: Résultats de l'exemple 4 en pourcentages des données du même type

Données… Type	**Positif**	**Négatif**	**Non classées**
Positives	45,02%	0%	54,98%
Négatives	11,94%	0%	88,06%

Tableau 3.12: Résultats de l'exemple 4, par groupes générés

	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	97,19%	---	85,14%
Prévalence	41,77%	0 %	58,23%

Cette combinaison plus permissive pour les homogénéités permet de générer plus de patrons qu'à l'exemple 1, où il n'y avait que 2 patrons positifs purs générés, mais également d'englober par erreur plus de données du type négatif. C'est pourquoi pour le groupe de type positif, le taux de présence est légèrement plus bas. On rassemble

toutefois presque 42% des données, alors qu'avec les l'homogénéité à 100% on ne regroupait même pas 15% des passagers.

On peut donc conclure que des différences aussi petites que de l'ordre de 1% au niveau des paramètres ont d'énormes répercussions sur les résultats. Pour augmenter le nombre de données classifiées, on peut soit diminuer la prévalence, soit rendre les homogénéités plus permissives. D'une manière ou de l'autre, les erreurs seront un peu plus nombreuses puisque l'on permettra à des patrons moins importants d'être utilisés dans le cas de la diminution de prévalence, ou dans le cas de l'homogénéité, à des patrons comportant plus d'erreurs. L'objectif demeure de trouver l'équilibre entre ces deux paramètres.

Une autre manière d'augmenter potentiellement la qualité des résultats consiste à utiliser des patrons plus longs, i.e. comprenant plus de termes. Toutefois, l'ajout d'un degré rallonge de beaucoup le temps de résolution: avec la première liste d'attributs, une résolution avec 3 degrés prenait environ 8 à 10 minutes, alors qu'une résolution avec 4 degrés prenait plus d'une heure et demie. Les résolutions à 5 degrés prenaient au-dessus de 6 heures! La nécessité de diminuer le nombre d'attributs est donc apparue évidente; c'est ce qui est décrit à la section suivante.

3.2 Épuration des attributs

Bien que l'on puisse se demander si une augmentation du nombre de degrés dans les patrons pourrait avoir un effet bénéfique sur la qualité des résultats, il n'est pas réaliste d'envisager de multiples séries d'essais à 4, 5 ou 6 degrés vu le temps de résolution trop élevé. Il a donc été décidé d'éliminer les attributs les moins utiles, afin d'avoir des délais de résolution plus acceptables et de pouvoir effectuer des essais sur le nombre de degrés.

En réalité, il n'y a pas seulement la suppression d'un attribut qui permet d'accélérer le temps de résolution: il y a aussi le fait d'avoir moins de valeurs différentes à l'intérieur d'un attribut. Par exemple, plutôt que d'avoir 26 classes de réservation (W, L, M, …), on peut les regrouper par produit (Tango, Tango plus,…). Ceci a le même effet que de diminuer le nombre d'attributs, car en fait, on diminue le nombre d'intervalles parmi lesquels on choisit pour construire des patrons. Il y a donc moins de combinaisons à explorer. En effet, le meilleur indicateur pour le temps de résolution est le nombre total de coupures: un attribut binaire équivaut à une coupure, tandis qu'un attribut contenant quatre valeurs différentes en contient trois. Le choix de ces coupures doit être fait en fonction du comportement des passagers. Il faut essayer de couper là où les changements apparaissent, mais pas trop souvent pour éviter d'avoir un grillage qui sépare chaque donnée seule à seule.

À l'aide d'Excel, on a calculé les taux de présence pour chaque valeur existant à l'intérieur de chaque attribut. Puis, si les taux de présence sont similaires pour chacune des valeurs, cela signifie que cet attribut n'a pas d'influence forte sur le taux de présence et qu'on peut l'éliminer. Ensuite, si les taux de présence sont très similaires pour certaines valeurs, mais pas pour toutes les autres, on regroupe les valeurs ayant des comportements similaires à l'intérieur de l'attribut.

Par exemple, le genre est un attribut comportant trois valeurs différentes (0, 1 et 2), on calcule les taux de présence pour tous les passagers ayant la valeur 0 (genre inconnu), puis le taux de présence pour ceux ayant la valeur 1 (hommes), puis pour la valeur 2 (femmes). Pour les valeurs 1 et 2, on observe que le taux diffère très peu, mais qu'il chute considérablement pour la valeur 0; on choisit donc de n'utiliser qu'un point de coupure pour cet attribut, entre le 0 et le 1. Tous les 2 sont remplacés par des 1. On a ainsi regroupé les valeurs 1 et 2 à l'intérieur de l'attribut « genre » pour éviter d'avoir un point de coupure superflu. De plus, cette manière de faire rend la phase de discrétisation

très simple et très rapide, puisqu'il suffit de placer des points de coupure entre chaque nouvelle valeur. On est ainsi assuré d'avoir séparé tous les changements de comportement des passagers avec un minimum de points.

Lorsque la génération des patrons est terminée et que le modèle est construit, le Datascope propose une analyse des attributs (bouton « Feature analysis ») qui affiche la proportion dans laquelle chaque attribut a été utilisé. Par exemple, si l'attribut apparaît dans cinq patrons positifs sur un total de dix, il obtient 0,5. On obtient donc un tableau comportant tous les attributs, ainsi que leur fréquence d'utilisation pour chacun des deux types (positif et négatif). Ceci permet de remarquer que certains attributs n'apparaissaient jamais (ou très peu) dans les positifs et/ou dans les négatifs. Les attributs les plus importants quant à eux apparaissent plus souvent dans au moins un des deux types de patrons. Pour les patrons positifs, on remarque une forte présence des attributs classe, OD_type, jour, heure, avance et groupe, tandis que pour les patrons négatifs, les attributs les plus utilisés sont la classe, la présence du numéro de billet, le point de vente, et l'origine. En retirant les attributs les moins significatifs (voir la liste des caractéristiques), on obtient la liste d'attributs finale présentée dans le tableau 2.1.

Évidemment, il est à signaler que ce choix est nécessaire pour accélérer les temps de résolution, mais en éliminant ainsi des attributs il se peut que l'on ait fait disparaître des patrons de degrés plus élevés qui auraient été générés ultérieurement, car les attributs retirés auraient peut-être servis à compléter un patron de degré 4, lorsqu'on aurait fait passer le nombre de degrés permis à 5. Ce risque est toutefois estimé comme très faible, et les temps de résolution ont considérablement diminué. Pour des résolutions avec 3 degrés, le temps passe de 8-10 minutes à 2-3 minutes, pour 4 degrés, le temps passe de 1,5 heures à 8-11 minutes, et pour 5 degrés de plus de 6 heures à 28-34 minutes.

Ultimement, puisque l'attribut dérivé de la classe apparaissait dans presque tous les patrons à tout coup, nous avons séparé les 38 501 passagers du mois de mars en cinq échantillons correspondant aux cinq grands ensembles de classes, appelés aussi les produits: Tango, Tango plus, Latitude, Affaires, Aéroplan. Ceci possède le double avantage d'éliminer un attribut comportant cinq valeurs, en plus d'avoir le même effet que d'ajouter un degré à tous les patrons, soit la classe. En revanche, il faudra faire cinq modèles LAD. Pour certains modèles, ceci permet également d'éliminer d'autres attributs. Par exemple pour le modèle Aéroplan, l'attribut voyageur fréquent est toujours égal à 0, car personne ne cumule des points dans cette catégorie qui désigne les passagers qui utilisent leurs points cumulés pour voyager. Pour le modèle Affaires, toutes les options sont inclues avec ce type de billet, donc l'attribut Go_options est inutile.

3.3 Nombre de degrés

Le nombre de degrés permis est une borne supérieure sur le nombre de termes (attributs) que peut contenir un patron. Lorsqu'on travaille avec 4 degrés, les patrons de degré 1, 2 et 3 sont inclus également, donc plus le degré est élevé, plus l'on peut s'attendre à obtenir un grand nombre de patrons. En revanche, pour qu'une donnée soit recouverte par un patron, elle doit satisfaire à chacune de ses conditions; à priori des patrons plus spécifiques devraient donc permettre de faire moins d'erreurs de classification, car les observations seraient similaires en plus de points.

Afin de déterminer quel est le nombre de degrés optimal à utiliser, il faut tenir compte du temps de résolution requis, de la quantité des patrons générés, mais aussi de leur qualité qui se mesure par les erreurs dans le classement et par la dégradation entre l'apprentissage et la validation. Après plusieurs essais en faisant varier le nombre de degrés de 1 jusqu'à 6, il apparaît que les patrons de degré 4 sont ceux qui offrent le meilleur équilibre et qui conviennent le mieux pour les expérimentations requises.

Il est facile de rejeter les options degrés 1 et 2, car très peu de patrons de degré aussi faible existent. En fait, ils existent, mais à des homogénéités très permissives. Ceci s'explique très aisément: supposons que parmi les passagers qui possèdent des billets électroniques, 94% sont venus. Pour qu'un patron ne comportant que cet attribut soit généré, il faudrait que l'homogénéité permise soit de 94% ou moins. C'est donc dire que pour créer des groupes qui ont des taux de présence le plus près de 100%, il est nécessaire de combiner plusieurs facteurs, car il n'en existe aucun qui ait une corrélation parfaite et directe avec la présence des passagers. Rappelons qu'il s'agit là d'une des principales forces de la méthode LAD car elle offre la possibilité de combiner les effets des attributs sur le comportement des passagers.

Les tests avec les degrés plus élevés, comme 5 ou 6, ont été moins convaincants que prévu. Il n'est pas clair que leur apport est significatif. D'abord ils nécessitent des temps de résolution qui sont très longs, allant de 30 minutes à plusieurs heures, et le nombre de patrons qui apparaît est faramineux, même que la plupart du temps, il est plus élevé que le nombre d'observations. Ceci est causé par l'algorithme utilisé pour générer les patrons dans le LAD Datascope, car il permet les redondances. Par exemple, si les quatre premiers attributs forment déjà un patron valide, on retient ce patron, en plus de lui ajouter par la suite différents attributs de manière à créer cinq ou six autres patrons, tous dérivés de celui-ci, mais de degré supérieur. Le phénomène observé est la présence de plusieurs patrons pour décrire une ou plusieurs mêmes données. Sans doute que l'utilisation d'un algorithme qui génère moins de redondances pourrait résorber une partie de ce problème. De plus, les résultats du classement ne sont pas significativement de meilleure qualité, ils demeurent essentiellement les mêmes à quelques données près.

L'équilibre entre le nombre de patrons et le temps de résolution se situe donc entre les degrés 3 et 4, qui prennent respectivement de 2 à 3 minutes, selon le nombre de données,

et de 8 à 11 minutes, sur le logiciel. Voici des tableaux qui présentent deux essais ayant chacun été réalisés pour 3 degrés et pour 4 degrés.

Essai 1

Homogénéité positive: 100% - négative: 15%
Prévalence: 8%

Tableau 3.13: Résultats de l'essai 1, avec 3 degrés, 24 patrons positifs

Apprentissage			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	100,00%	---	87,18%
Prévalence	23,39%	0 %	76,61%
Validation			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	93,86%	---	88,93%
Prévalence	25,09%	0 %	74,76%
Ensemble des observations			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	96,82%	---	88,04%
Prévalence	24,24%	0 %	75,76%

Tableau 3.14: Résultats de l'essai 1, avec 4 degrés, 282 patrons positifs

Apprentissage			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	100,00%	---	82,97%
Prévalence	42,30%	0 %	57,70%
Validation			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	95,35%	---	86,07%
Prévalence	44,17%	0 %	55,83%
Ensemble des observations			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	97,63%	---	84,50%
Prévalence	43,23%	0 %	56,77%

Essai 2

Homogénéité positive: 99% - négative: 20%

Prévalence: 10%

Tableau 3.15: Résultats de l'essai 2, avec 3 degrés, 24 patrons positifs

Apprentissage			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	98,66%	---	86,98%
Prévalence	27,35%	0 %	72,65%
Validation			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	93,23%	---	88,97%
Prévalence	28,17%	0 %	71,83%
Ensemble des observations			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	95,90%	---	87,97%
Prévalence	27,76%	0 %	72,24%

Tableau 3.16: Résultats de l'essai 2, avec 4 degrés

Apprentissage			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	97,61%	---	83,02%
Prévalence	49,05%	0 %	50,95%
Validation			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	95,22%	---	85,30%
Prévalence	49,08%	0 %	50,92%
Ensemble des observations			
	Groupe Positif	**Groupe Négatif**	**Groupe Non classés**
Homogénéité	96,41%	---	84,16%
Prévalence	49,06%	0 %	50,94%

Pour chacun des deux essais, on remarque que le nombre de patrons créés lorsque le degré 4 est permis est dix fois plus élevé que le nombre de patrons de degré 3 seulement. Tandis que les patrons de degré 3 ne réussissent à classifier que 27,76% des données, les

patrons de degré 4 peuvent en classifier près de la moitié, 49,06%, soit presque deux fois plus. Un deuxième phénomène rend les patrons de degré 4 beaucoup plus intéressants que ceux de degré 3: lorsqu'on examine plus spécifiquement les taux de présence (homogénéités) des deux parties de la base de données, apprentissage et validation, on s'aperçoit immédiatement que lorsqu'on utilise 4 degrés, il y a moins d'écarts entre les deux groupes. Ceci est très important car la première partie de la base de données sert à la création; il est donc normal que les résultats y soient meilleurs. Quant à la deuxième, il s'agit de la validation des patrons sur un ensemble de données différentes, donc il est souhaitable de voir les résultats se dégrader le moins possible du premier tableau au deuxième, et aussi d'y observer des homogénéités qui sont semblables.

Non seulement y a-t-il moins d'écarts entre les deux groupes lorsque l'on travaille avec 4 degrés, mais de plus, si on regarde le total des observations, les essais réalisés avec 4 degrés comportent des taux de présence plus près de 100% que ceux du degré 3 (96,41% vs 95,90%). Donc avec 4 degrés, on classifie plus de données, les résultats se comportent de manière plus stable lors du passage de la phase d'apprentissage à la validation et le taux de présence est plus près de ce que l'on cherche à obtenir. C'est pour l'ensemble de ces raisons que le nombre de degrés a été fixé à 4 pour les expérimentations subséquentes.

3.4 Autres tests

Il y a certains processus qui ont été imaginés et testés durant le projet, mais qui n'ont toutefois pas démontré suffisamment leur supériorité ou valeur ajoutée pour être retenus. Certaines de ces idées, explorées plus en profondeur ou appliquées différemment, pourraient potentiellement améliorer les résultats.

Il a été observé que lorsque l'on met des paramètres trop stricts, aucun patron n'est créé. À mesure que l'on élargi les critères, de plus en plus de patrons sont générés. Il se peut par exemple qu'avec une prévalence de 12%, aucun patron n'ait une homogénéité de 100%, mais qu'en n'exigeant seulement une prévalence de 4% des centaines de patrons existent. Parmi ceux-ci, il pourrait y en avoir qui auraient des prévalences de 5%, mais aussi de 10%. Comme l'objectif premier demeure de faire le moins d'erreurs de classement possible, et que les patrons qui ont les prévalences les plus élevés (pour une même homogénéité) sont meilleurs, nous avons imaginé un processus itératif qui constituait à retirer les données classifiées et à refaire la méthode LAD pour les données non-classifiées restantes. Cette idée était fondée sur le fait qu'il existe peut-être des patrons parmi les données non-classifiées, mais qu'ils ne sont pas suffisamment importants pour apparaître avec les prévalences exigées. Par contre, une fois que les données correspondant à un premier ensemble de patrons sont dégagées de l'ensemble, la prévalence de 5% peut être plus facilement réalisable. Aussi, comme un certain nombre de données seront retirées, il serait peut-être plus facile de créer des ensembles de conditions pour lesquels les passagers sont tous présents (ou absents).

Plusieurs interrogations naissent de cette idée quant à la manière de faire. Combien de patrons choisir, comment fixer les paramètres d'une itération à l'autre, etc. Un peu à tâtons, nous avons bien vite réalisé que le processus manquait de rigueur. Nous avons donc révisé ce processus itératif et fait quelques ajustements. Il s'agissait maintenant de commencer avec des paramètres stricts, et dès l'apparition d'un premier patron, retirer les données classées puis reprendre. Ce processus s'est avéré plutôt long, et les résultats décevants. Effectivement, le fait de ne choisir qu'un seul patron à la fois faisait en sorte que beaucoup plus d'itérations étaient nécessaire avant d'en arriver essentiellement aux mêmes résultats obtenus après 2 ou 3 itérations du premier processus. De plus, les données restantes étaient peu séparables et très peu de patrons pouvaient en être dégagés. Nous avons donc mis de côté l'idée du processus itératif.

Il arrive parfois que des données satisfassent à des patrons des deux types (positif et négatif) à la fois, surtout lorsque le nombre de patrons est élevé. Le logiciel propose une manière de faire qui consiste à calculer la proportion des patrons satisfaits de chaque type, et de prendre celle qui est la plus prépondérante. Nous nous sommes questionnés quant à savoir si ne pas classer ces données du tout pourrait améliorer les résultats. Nous avons également essayé en utilisant la somme des prévalences absolues correspondant à chaque patron satisfait. Malgré tout, le logiciel propose tout de même la meilleure manière de procéder. Il serait intéressant de fixer des seuils critiques différents de 0 pour le classement final. Ainsi, les données qui sont légèrement plus positives, ou à peine plus négatives que 0 ne seraient pas classifiées.

À la fin de cette phase d'exploration, on peut donc conclure plusieurs choses: la meilleure séparation entre l'apprentissage et la validation réside dans la répartition 50%-50%, certains attributs peuvent être enlevés sans impact majeur, certaines valeurs à l'intérieur des attributs peuvent aussi être regroupées, les patrons de degré 4 offrent de très bons résultats et peuvent être générés en un temps de résolution convenable. Cette phase exploratoire a également permis de relever quelques lacunes et limites de la méthode qui sont élaborées à la troisième section du chapitre 4.

CHAPITRE 4 RÉSULTATS ET DISCUSSION

Malgré le fait qu'il se soit avéré moins aisé que prévu de réellement séparer les passagers en deux groupes distincts (les présents et les absents), les résultats s'avèrent très encourageants au bout du compte. Pour chacun des cinq produits, on applique les patrons générés à partir des données du mois de mars 2009 par la méthode LAD sur les données du mois d'avril 2009. Pour chaque passager, on indique s'il sera présent, absent ou incertain selon les patrons auxquels il satisfait les conditions. Ensuite, pour chacun des trois groupes qui sont formés, on applique le taux de présence du mois de mars correspondant, i.e. l'homogénéité du groupe. Ainsi, on obtient le nombre de passagers prévus pour avril. Les passagers peuvent ainsi être regroupés par vols, puis par cabine, par produit et finalement par classe.

Ce chapitre se divise en trois sections: d'abord les résultats finaux globaux pour les cinq modèles seront présentés, ensuite nous comparons la méthode LAD avec PROS, l'outil de prévision commercial utilisé actuellement chez Air Canada pour estimer le nombre de passagers présents et absents, et finalement, la dernière section regroupe un ensemble de suggestions qui pourraient potentiellement améliorer les résultats.

4.1 Résultats finaux

À l'aide des données extraites du mois de mars 2009, la méthode LAD a été exécutée cinq fois, une pour chaque produit: Tango, Tango plus, Latitude, Affaires et Aéroplan. Les résultats retenus sont présentés ci-dessous.

Tango

Homogénéité positive: 99% - négative: 15%

Prévalence: 3%

Nombre de patrons générés: 315 positifs, 485 négatifs

Tableau 4.1: Résultats Tango (classes T, E, P, G, N, K, R)

	Groupe positif	**Groupe négatif**	**Groupe non classés**
Homogénéité	98,16%	22,22%	93,13%
Prévalence	23,62%	0,61%	75,78%

Tango plus

Homogénéité positive: 98% - négative: 15%

Prévalence: 3%

Nombre de patrons générés: 358 positifs, 475 négatifs

Tableau 4.2: Résultats Tango plus (classes B, H, V, Q, A, L, S)

	Groupe positif	**Groupe négatif**	**Groupe non classés**
Homogénéité	97,02%	19,59%	86,46%
Prévalence	25,24%	3,01%	71,75%

Latitude

Homogénéité positive: 99% - négative: 15%

Prévalence: 3%

Nombre de patrons générés: 78 positifs, 236 négatifs

Tableau 4.3: Résultats Latitude (classes Y, M, U)

	Groupe positif	**Groupe négatif**	**Groupe non classés**
Homogénéité	94,39%	19,63%	73,91%
Prévalence	16,67%	4,35%	78,98%

Affaires

Homogénéité positive: 99% - négative: 15%

Prévalence: 3%

Nombre de patrons générés: 160 positifs, 646 négatifs

Tableau 4.4: Résultats Affaires (classes J, C, Z, I)

	Groupe positif	**Groupe négatif**	**Groupe non classés**
Homogénéité	95,19%	36,67%	81,64%
Prévalence	34,15%	3,52%	62,32%

Aéroplan

Homogénéité positive: 100% - négative: 40%

Prévalence: 5%

Nombre de patrons générés: 580 positifs, 667 négatifs

Tableau 4.5: Résultats Aéroplan (classes W et D)

	Groupe positif	**Groupe négatif**	**Groupe non classés**
Homogénéité	97,88%	41,67%	83,22%
Prévalence	53,65%	2,20%	44,15%

Dans les tableaux présentés ci-dessus, l'homogénéité du groupe est aussi le taux de présence de ce même groupe. C'est ce taux que nous utilisons pour le calcul du nombre total de personnes présentes avec la méthode LAD. La manière de prédire le nombre de personnes qui seront présentes pour leur vol consiste donc à classifier d'abord tous les nouveaux passagers dans les trois groupes, puis à multiplier le taux de présence prévu pour chaque groupe avec le nombre de passagers placés dans ce groupe. On fait la somme des présences absolues. Si l'on divise cette somme par le nombre de réservations effectuées, on obtient également le taux de présence prévu.

Les PNRs et leurs caractéristiques extraits pour mois d'avril 2009 ont d'abord été discrétisés de la même manière que les observations du mois de mars, afin de pouvoir les comparer aux patrons. Puis les données ont été séparées selon les cinq produits. Par la suite, elles ont été comparées aux patrons de mars 2009 de leur produit respectif. En fonction du nombre de patrons positifs et négatifs auxquels l'observation (passager) correspond, on calcule le score de chaque observation. On lui assigne une présence (1), si le score est plus grand que 0, une absence (0) s'il est en-dessous de 0, ou la mention d'incertitude (9) si le score est nul. La manière d'obtenir le nombre de présences prévues par la méthode LAD est donc de multiplier la colonne du nombre de valeurs comptées pour chaque groupe avec le taux de présence du groupe, et de faire la somme pour les trois catégories de passagers enregistrés.

Tableau 4.6: Résultats détaillés de la méthode LAD pour Tango, avril 2009

Tango	**Valeur**	**Nombre**	**Taux**
Présents	1	3 219	98,16%
Absents	0	58	22,22%
Incertains	9	10 083	93,13%

Selon le tableau ci-dessus, parmi tous les passagers voyageant de Vancouver à Calgary, durant le mois d'avril 2009, dans une des classes du produit Tango, 3 219 satisfont plutôt des patrons positifs, 58 des patrons négatifs, et 10 083 aucun. Les taux de présence sont ceux obtenus dans les groupes équivalents pour le mois de mars.

Tableau 4.7: Comparaison de la méthode LAD avec le taux réel pour Tango, avril 2009

Tango	**Réel**	**LAD**
Présents	12 575	12 563
Total	13 360	13 360
Taux	94,03%	94,12%

Si pour chacun des trois groupes on multiplie le nombre de passagers par le taux de présence associé, et qu'on en fait la somme, on obtient 12 563 présences prévues parmi les 13 360 passagers. En réalité, il y en a 12 575 qui sont venus.

Voici également la même comparaison pour les quatre autres produits.

Tableau 4.8: Comparaison de la méthode LAD avec le taux réel pour Tango plus, avril 2009

Tango plus	**Réel**	**LAD**
Présents	15 783	16 037
Total	18 384	18 384
Taux	85,85%	87,24%

Tableau 4.9: Comparaison de la méthode LAD avec le taux réel pour Latitude, avril 2009

Latitude	**Réel**	**LAD**
Présents	2165	2179
Total	2876	2876
Taux	75,28%	75,76%

Tableau 4.10: Comparaison de la méthode LAD avec le taux réel pour Affaires, avril 2009

Affaires	**Réel**	**LAD**
Présents	1476	1462
Total	1701	1701
Taux	86,77%	85,93%

Tableau 4.11: Comparaison de la méthode LAD avec le taux réel pour Aéroplan, avril 2009

Aéroplan	**Réel**	**LAD**
Présents	3026	3009
Total	3365	3365
Taux	89,38%	89,52%

Bien que cela ne soit pas suffisant pour démontrer l'efficacité de la méthode, on constate toutefois que pour chacun des cinq produits, le taux de présence (tous les vols confondus durant le mois d'avril 2009) est presque identique au taux réel de présence. Ceci est un premier indicateur que la méthode semble fonctionner globalement. Toutefois, il est important de grouper les passagers par vols pour calculer les prévisions. En effet, c'est la capacité physique de l'avion qui limite le nombre de passagers ou qui détermine le nombre de places vacantes.

Puisque les prévisions pour chaque passager du mois d'avril ont été établies, il est possible de calculer les taux de présence pour un vol. De la même manière, on pourrait aussi les calculer pour une cabine, pour un produit et/ou pour une classe, selon ce qui nous intéresse. Il suffit simplement de regrouper les passagers pour lesquels on désire faire le calcul, compter les 0, les 1 et les 9, multiplier par les taux correspondants déjà calculés à l'aide des modèles du mois de mars (et qui varient selon le produit), et le nombre obtenu est le nombre de passagers prévus. C'est ce qui est fait à la section suivante.

4.2 Comparaison avec PROS

Comme mentionné précédemment, PROS est l'outil commercial utilisé chez Air Canada pour effectuer les prévisions de ventes. PROS fournit des prévisions différentes pour chaque jour de la semaine, par vol, par classe. Pour avoir une comparaison la plus

objective possible, nous avons choisi de prendre les données de PROS pour le mois de mars 2009 seulement. En pratique, PROS peut utiliser un historique variant de quatre semaines jusqu'à deux ans. On utilise chez Air Canada généralement un an.

À l'aide de toutes les données historiques récupérées via PROS, on calcule donc un taux de présence par classe, par vol, par jour. On applique par la suite ce taux sur le nombre de réservations effectuées pour obtenir le nombre de passagers que PROS prévoit qui seront présents.

Comme PROS émet de prévisions par jour, par vol et par classe, nous avons également séparé toutes les données du mois d'avril de cette manière. Par souci de simplicité, nous avons retenu quatre vols différents (202, 210, 214, 224), évalués lors de quatre journées: le 1er avril, le 5 avril, le 10 avril et le 13 avril. Comme il existe 23 classes, ces quatre vols durant ces quatre jours ont généré 368 points de comparaison. Pour ces 368 points, nous avons calculé les coefficients de relation entre les deux méthodes et la réalité. On remarque dans le tableau suivant que les prévisions effectuées par la méthode LAD sont plus près de la réalité que celles de PROS.

Tableau 4.12: Coefficients de corrélations entre les présences réelles et celles prévues par la méthode LAD et PROS

	Présences réelles	**Prévisions LAD**	**Prévisions PROS**
Présences réelles	1		
Prévisions LAD	0.99635	1	
Prévisions PROS	0.97118	0.96371	1

Dû à la manière dont les billets sont mis en vente chez Air Canada (allocation en « nesting ») il est souhaitable d'analyser les prévisions par classe. Par contre, sur un vol,

lors d'une journée donnée, pour une classe donnée, il n'y a en moyenne que quatre ou cinq passagers. Vu l'impossibilité de comparer des valeurs aussi petites, nous devons, comme mentionné précédemment, agréger ces données (les passagers de chaque classe) soit par produit, soit par cabine, soit par vol, pour avoir des informations plus significatives. Comme les tarifs varient d'une classe à l'autre à l'intérieur d'un même produit, en agglomérant les passagers de plusieurs classes ensemble, nous ne pourrons estimer réellement les profits générés.

Le taux de présence par vol est le meilleur indicateur pour la survente si l'on ne tient compte que des sièges physiques dans l'appareil. Si l'on veut mieux prendre en compte les tarifs, on peut considérer le plus grand écart de tarif entre les billets. Il se trouve entre la cabine J, en avant de l'appareil, et la cabine Y, à l'arrière. Les résultats sont donc présentés par vol et par cabine dans les tableaux suivants. Le même travail a également été réalisé pour chacun des produits (Annexe 2).

Les indicateurs de qualité utilisés pour les résultats sont le carré des erreurs, les erreurs relatives, pour lesquelles on peut calculer la somme, la moyenne et l'écart type, ainsi que le nombre de fois que la méthode LAD a été meilleure que PROS et le taux de succès équivalent. Voici donc le tableau qui regroupe chacun des 16 vols reconstitués à l'aide des fichiers disponibles à l'aéroport après le départ.

Tableau 4.13: Résultats comparatifs pour les 16 vols évalués

		Prévisions		Mesures de l'erreur LAD		Mesures de l'erreur PROS		Meilleur
Vol	**Présences réelles**	**Prévision LAD**	**Prévision PROS**	**Err.^2**	**Err. Rel.**	**Err.^2**	**Err. Rel.**	**PROS/ LAD**
1	47	75.916	81.405	836.147	61.5%	1183.698	73.2%	L
2	59	60.574	60.203	2.478	2.7%	1.447	2.0%	P
3	70	69.234	76.332	0.587	1.1%	40.098	9.0%	L
4	110	104.588	113.073	29.288	4.9%	9.445	2.8%	P
5	85	79.187	94.704	33.793	6.8%	94.161	11.4%	L
6	121	120.634	125.072	0.134	0.3%	16.581	3.4%	L
7	105	105.551	118.163	0.303	0.5%	173.275	12.5%	L
8	91	91.025	92.106	0.001	0.0%	1.223	1.2%	L
9	107	107.143	95.287	0.020	0.1%	137.194	10.9%	L
10	114	112.664	114.635	1.784	1.2%	0.404	0.6%	P
11	72	68.096	63.849	15.238	5.4%	66.444	11.3%	L
12	44	53.466	71.361	89.599	21.5%	748.616	62.2%	L
13	129	123.103	127.760	34.772	4.6%	1.539	1.0%	P
14	122	118.295	133.100	13.727	3.0%	123.220	9.1%	L
15	96	94.180	87.009	3.314	1.9%	80.833	9.4%	L
16	111	111.521	120.747	0.272	0.5%	94.996	8.8%	L
	Somme			1061.457	54.6%	2773.175	228.8%	
	Moyenne			66.341	3.6%	173.323	14.3%	
	Écart type			206.624	5.4%	324.115	21.4%	
							Nbre P	4
							Nbre L	12
							LAD:	75.00%

Il est difficile de comparer la méthode LAD et les prévisions de PROS car le principe de base n'est pas du tout le même: le premier classifie les passagers selon des caractéristiques qui leur sont propres, ainsi que du risque de présence, tandis que le deuxième se base uniquement sur un pourcentage historique et l'applique à un groupe. Toutefois, il est possible d'affirmer que la méthode LAD propose des prévisions qui s'avèrent plus fidèles à la réalité que celles de PROS, du moins pour les vols 202, 210, 214 et 224 des 1er, 5, 10 et 13 avril.

Ceci n'est pas suffisant pour établir que la méthode LAD génère théoriquement plus de profits que PROS. En effet, puisque les billets sont mis en vente à l'avance, en « nesting » (allocation protégée) et à des tarifs différents pour chaque classe, il est très difficile d'établir les profits associés à la politique de survente uniquement. Il ne s'agit là que d'une petite partie du processus de distribution des passagers à travers les classes.

Un passager assis dans la cabine J (à l'avant) peut payer jusqu'à trois ou quatre fois le prix du passager installé dans la cabine Y (à l'arrière). Nous avons donc, pour chacun des 16 vols, séparé les classes des deux cabines.

Tableau 4.14: Résultats comparatifs pour les 16 vols évalués, cabine J

			Prévisions		Mesures de l'erreur LAD		Mesures de l'erreur PROS		Meilleur
VOL	Date	Présences réelles	Prévision LAD	Prévision PROS	Err.^2	Err. Rel.	Err.^2	Err. Rel.	PROS/ LAD
202	1	0	0.000	0.000	0.000		0.000		L
210	1	4	3.433	3.514	0.322	14.2%	0.236	12.1%	P
214	1	7	5.882	7.929	1.250	16.0%	0.862	13.3%	P
224	1	5	6.044	5.800	1.091	20.9%	0.640	16.0%	P
202	5	8	7.017	9.591	0.967	12.3%	2.531	19.9%	L
210	5	2	1.811	1.286	0.036	9.5%	0.510	35.7%	L
214	5	5	6.017	7.683	1.035	20.3%	7.196	53.7%	L
224	5	6	5.212	4.673	0.621	13.1%	1.762	22.1%	L
202	10	4	3.861	3.822	0.019	3.5%	0.032	4.4%	L
210	10	8	8.116	4.225	0.014	1.5%	14.251	47.2%	L
214	10	4	3.742	2.393	0.067	6.5%	2.583	40.2%	L
224	10	2	4.260	3.917	5.109	113.0%	3.674	95.8%	P
202	13	8	7.371	7.594	0.396	7.9%	0.165	5.1%	P
210	13	3	3.795	15.083	0.632	26.5%	146.007	402.8%	L
214	13	4	3.297	3.048	0.494	17.6%	0.905	23.8%	L
224	13	6	5.337	5.750	0.440	11.1%	0.063	4.2%	P
			Somme		12.490	293.6%	181.415	796.2%	
			Moyenne		0.781	19.6%	11.338	53.1%	
			Ecart type		1.227	26.7%	36.100	99.7%	
								Nombre P	6
								Nombre L	10
								Taux LAD	62.50%

Tableau 4.15: Résultats comparatifs pour les 16 vols évalués, cabine Y

			Prévisions		Mesures de l'erreur LAD		Mesures de l'erreur PROS		Meilleur
VOL	Date	Présences réelles	Prévision LAD	Prévision PROS	Err.^2	Err. Rel.	Err.^2	Err. Rel.	PROS/ LAD
202	1	47	75.916	81.405	836.147	61.5%	1183.698	73.2%	L
210	1	55	57.141	56.688	4.586	3.9%	2.851	3.1%	P
214	1	63	63.352	68.404	0.124	0.6%	29.200	8.6%	L
224	1	105	98.544	107.273	41.681	6.1%	5.168	2.2%	P
202	5	77	72.170	85.113	23.328	6.3%	65.817	10.5%	L
210	5	119	118.823	123.786	0.031	0.1%	22.908	4.0%	L
214	5	100	99.534	110.481	0.218	0.5%	109.848	10.5%	L
224	5	85	85.813	87.433	0.660	1.0%	5.920	2.9%	L
202	10	103	103.281	91.465	0.079	0.3%	133.061	11.2%	L
210	10	106	104.548	110.410	2.108	1.4%	19.451	4.2%	L
214	10	68	64.355	61.456	13.288	5.4%	42.826	9.6%	L
224	10	42	49.206	67.444	51.919	17.2%	647.407	60.6%	L
202	13	121	115.732	120.165	27.751	4.4%	0.697	0.7%	P
210	13	119	114.500	118.017	20.247	3.8%	0.966	0.8%	P
214	13	92	90.882	83.961	1.249	1.2%	64.629	8.7%	L
224	13	105	106.185	114.997	1.404	1.1%	99.932	9.5%	L
			Somme		1024.821	114.6%	2434.380	220.3%	
			Moyenne		64.051	7.2%	152.149	13.8%	
			Ecart type		206.537	15.1%	316.419	21.2%	
								Nombre P	4
								Nombre L	12
								Taux LAD	75.00%

En examinant les tableaux de plus près, on remarque que la méthode LAD occasionne des erreurs plus petites que PROS (somme, moyenne et écart type), que ce soit pour les erreurs au carré ou pour les erreurs relatives, et ce pour les vols complets, pour la cabine J, et pour la cabine Y. Il en va de même lorsque l'on analyse les données par produit (voir Annexe 2).

Le meilleur indicateur dans un cas comme celui-ci est la somme des erreurs au carré. La méthode LAD, pour la cabine Y obtient la somme de 1024,821, tandis que PROS obtient 2434,380. L'erreur moyenne de la méthode LAD est moins de la moitié de celle de PROS, et son écart type moins des deux tiers. De plus, dans 75% des vols étudiés, la méthode LAD est plus près de la réalité que PROS. Pour la cabine J, les résultats obtenus par la méthode LAD sont tout aussi remarquables, mais cette cabine comporte peu de passagers et présente beaucoup de variabilité, donc ces statistiques doivent être interprétées avec un peu plus de vigilance.

Bien entendu, nous n'avons évalué que quatre vols, lors de quatre journées, en utilisant les données du mois de mars, pour prédire avril. Il reste encore plusieurs choses à faire avant de remplacer PROS par la méthode LAD pour effectuer les prévisions de ventes. En plus d'examiner d'autres vols à d'autres moments, il serait intéressant d'appliquer la méthode à une autre paire de villes. On pourrait ainsi observer dans quelle mesure les patrons seraient différents, et aussi s'assurer que les résultats sont tout aussi bons. Il serait ainsi possible de déterminer si la méthode LAD présente des faiblesses sur certains marchés, ou encore, si elle est systématiquement meilleure que PROS à tous les coups.

Pour le moment, on peut affirmer que sur le marché Vancouver-Calgary, pour le mois d'avril 2009, la méthode LAD se rapproche significativement plus de la réalité que les prévisions de PROS.

4.3 Améliorations potentielles

Plusieurs améliorations sont encore possibles et devraient être apportées. Cette section comporte des suggestions regroupées en deux: celles qui touchent les étapes préliminaires ou parallèles à la méthode LAD et les variations à l'intérieur de la méthode LAD elle-même.

Puisque le logiciel limite le nombre de PNR, la méthode LAD a été appliquée pour un mois de données. Pour être conséquents, il n'a fallu utiliser qu'un mois de données de PROS pour la comparaison. Il serait donc fortement souhaitable de reprendre la méthode avec plus de PNR, par exemple tous ceux de l'année 2009. Par la suite, il serait possible de comparer ces résultats avec un an de données historiques dans PROS. De plus, pour mieux modéliser la réalité en fonction du temps, il est suggéré de n'utiliser que les PNR enregistrés jusqu'à une certaine date limite, par exemple 10 jours avant le vol, 5 jours avant le vol, 1 jour avant le vol, afin de faire des projections de vente et ajuster la survente en temps réel. On pourrait ainsi établir si la méthode LAD performe mieux que PROS, même lorsqu'il n'y a que peu de réservations faites sur un vol, pour prédire les ventes finales.

Des contraintes liées aux temps de résolution nous ont également obligés à limiter le nombre d'attributs retenus à ceux que nous savions déjà suffisamment pertinents. Par contre, si on utilise la méthode LAD à son plein potentiel, on pourrait utiliser de nombreux autres attributs, et c'est lors de la création des patrons que nous pourrions juger de ceux qui sont superflus. En effet, la méthode LAD peut permettre d'identifier un ensemble minimal de variables à l'intérieur d'un ensemble beaucoup plus vaste, à partir desquelles on peut générer tous les patrons nécessaires à la classification. Dans la littérature on fait référence à ce sous-ensemble de variables comme étant le « *minimal support set* ».

Un autre aspect important de la méthode LAD que nous n'avons que très peu abordé est le système des points de coupures. Cette partie a été réalisée essentiellement à l'aide d'Excel et aussi des suggestions des employés d'Air Canada. Par contre, la méthode LAD propose des manières d'élaborer des systèmes de points de coupures qui mériteraient d'être explorées plus en profondeur.

Dans cette étude, nous n'avons pas traité l'aspect de la génération des patrons, car ceci était fait directement par le logiciel. Toutefois, il existe plusieurs manières d'exécuter cette partie de la méthode. Une très grande partie de la littérature au sujet de la méthode LAD traite des différents algorithmes, ayant chacun leurs forces et leurs faiblesses respectives. La plupart des approches sont dérivées d'une des trois manières plus générales: « *top-down* » qui consiste à prendre chaque observation et à en faire un patron caractéristique, puis à jumeler les patrons en les raccourcissant, à l'opposé, l'approche « *bottom-up* » débute avec des patrons de degré 1 et procède en les rallongeant, quant à la troisième manière de faire, il s'agit des méthodes hybrides.

La méthode LAD comporte une phase à la fin qui consiste à sélectionner les patrons parmi ceux générés qui seront retenus pour former le modèle. Pour nos expérimentations, nous avons simplement conservé l'ensemble des patrons créés, mais il est important de souligner qu'il pourrait en être différemment. À cette étape-ci, il suffit résoudre un problème de recouvrement pour limiter le nombre de patrons dans le modèle. On peut ainsi en générer un plus grand nombre, sachant que l'on ne choisira que les meilleurs, augmentant ainsi la qualité du modèle.

Finalement, à la toute fin de la méthode, on classifie les observations selon qu'elles ont des scores positifs, négatifs ou nuls. En théorie, il est possible de fixer des seuils critiques différents de zéro. Par exemple, pour qu'une donnée soit classifiée positive, il faudrait que son score soit au moins de 0,5. Ceci occasionnerait plus de données non-

classifiées, mais potentiellement moins d'erreurs classifications aussi. C'est donc une avenue qui vaudrait la peine d'être étudiée.

Donc il y a essentiellement deux séries d'améliorations à porter à ce modèle: premièrement raffiner les résultats avec la méthode telle qu'utilisée jusqu'à présent (autres vols, autres marchés), deuxièmement peaufiner la méthode LAD dont les forces ne sont pas encore exploitées au maximum (points de coupure, choix des patrons).

CONCLUSION

La méthode LAD a été étudiée et adaptée au problème de la survente des sièges en transport aérien afin de déterminer si cette nouvelle approche pouvait s'avérer utile pour les prévisions de surventes de sièges. En conclusion, cette méthode permet de classifier les passagers en trois groupes, ayant chacun leur taux de présence prévisionnel respectif. Puis, lorsqu'on en fait la somme pondérée, on obtient le nombre de passagers prévus pour le vol étudié.

Rappelons également que la particularité de ce modèle réside dans le fait que chaque passager est caractérisé par un vecteur d'attributs basés sur ses propres caractéristiques: classe de réservation, heure et le jour de départ, billet électronique, etc. C'est en posant des conditions sur certains de ces attributs que l'on forme des sous-ensembles de passagers étant plus prédisposés à être présents ou absents. Il a été validé que cette manière de faire correspond plus au comportement réel des passagers que l'utilisation simple d'un pourcentage de présence historique, approche actuellement en vigueur chez Air Canada (PROS). Pour le moment, seulement le vol Vancouver-Calgary a été évalué, mais il serait intéressant de développer le même modèle pour d'autres paires de villes.

Il est à noter qu'afin de simplifier le projet, nous n'avons tenu compte que des données disponibles du mois de mars 2009 pour les deux méthodes, et les avons appliquées sur les PNR du mois d'avril 2009. Cette comparaison avec le système actuel de prévision des ventes utilisé chez Air Canada (PROS) s'est révélée en faveur de la méthode LAD, qui s'est montrée très concurrentielle vis-à-vis PROS, mais avant d'adopter cette méthode il est nécessaire de la réévaluer avec davantage de données.

Au terme de ce projet, nous pouvons dégager plusieurs recommandations relativement à l'application de la méthode LAD sur des passagers en transport aérien: la séparation

apprentissage/validation doit être adéquate afin de permettre la création de patrons représentatifs des deux parties de l'échantillon, par exemple 50%-50%; les points de coupures doivent être mis en place de manière à permettre le plus de séparations possibles entre les données positives et négatives, mais également de telle sorte à ce qu'il y ait suffisamment d'observations dans chacune des divisions de la grille; les patrons de degré 4 peuvent être générés en un temps acceptable et offrent un classement tout aussi bon que les modèles utilisant 5 ou 6 degrés.

Toutefois, certaines lacunes ont été relevées: le logiciel étant limité à 16 000 observations, il serait nécessaire de programmer la génération de patrons afin de réaliser des tests comportant plus de données, ou de pouvoir tester différents algorithmes de génération, la qualité des données à l'entrée pourrait certainement être améliorée, soit par un meilleur choix d'attributs, soit par une meilleure compréhension de ceux-ci. En effet, les bases de données industrielles, en particulier en transport aérien, sont volumineuses et il manque parfois des paramètres pour certaines observations. La préparation des données représente un travail colossal et non trivial.

Pour terminer, de nouvelles voies de recherche sont proposées. En plus d'améliorer la phase du prétraitement des données et d'employer la méthode LAD à son plein potentiel, il serait intéressant de généraliser le modèle pour ne pas avoir à refaire la méthode LAD du début à la fin pour chaque paire de villes. Pour ce faire, il est possible de rajouter deux attributs, soit un pour l'origine du vol et l'autre pour sa destination. Une autre manière ne nécessitant pas autant de données consisterait à développer la méthode LAD pour les vols. Les observations ne seraient plus les passagers, mais plutôt les vols eux-mêmes. Les attributs comporteraient donc, par exemple, l'heure et la date de départ, le jour de la semaine, la proportion de passager arrivant d'une connexion, l'historique d'absentéisme sur ce vol. Ceci permettrait d'utiliser les données historiques enregistrées dans PROS comme un « input » à la méthode LAD.

Encore une fois, il est à noter que cette étude a permis d'identifier la méthode LAD comme étant très prometteuse pour effectuer de meilleures prévisions pour la survente des sièges, mais il s'agit d'une exploration seulement et la méthode n'a pas encore été utilisée à son plein potentiel. Ceci nous porte à croire que le travail mérite d'être fait compte tenu que des millions de dollars sont en jeu. Cette méthode a beaucoup de succès dans des domaines très diversifiés comme la médecine, la biologie, les finances, la psychologie, et maintenant, nous pouvons aussi dire en transport aérien.

BIBLIOGRAPHIE

Abramson, S. D., Alexe, G., Hammer, P. L., Kohn, J. (2005) A computational approach to predicting cell growth on polymeric biomaterials. *Wiley InterScience.* Consulté le 25 septembre 2008, tiré de http://dx.doi.org/10.1002/jbm.a.30266

Agard, B., Kusiak, A. (2005). Exploration des bases de données industrielles à l'aide du data mining – Perspectives. *9e Colloque National AIP-PRIMECA*, La Plagne, France.

Alexe, G., Alexe, S., Axelrod, D., Hammer, P. L., Weissman, D. (2005). Logical analysis of diffuse large B-cell lymphomas. *Artificial Intelligence in Medicine, 34*, 235-267.

Alexe, G., Hammer, P. L. (2006a). Spanned patterns for the logical analysis of data. *Discrete Applied Mathematics, 154* (7), 1039-1049.

Alexe, G., Alexe, S., Hammer, P. L. (2006b). Pattern-based clustering and attribute analysis. *Soft Comput, 10* (5), 442-452.

Alexe, S., Blackstone, E., Hammer, P. L. (2003). Coronary risk prediction by logical analysis of data. *Annals of Operations Research, 119*, 15-42.

Alexe, S., Hammer, P. L. (2006). Accelerated algorithm for pattern detection in logical analysis of data. *Discrete Applied Mathematics, 154* (7), 1050-1063.

Bennane, A., Yacout, S. (2009). LAD-CBM; new data processing tool for diagnosis and prognosis in condition-based maintenance. *J Intell Manuf.* Consulté le 7 janvier 2010, tire de http://dx.doi.org/10.1007/s10845-009-0349-8

Bonates, T. O., Hammer, P. L., Kogan, A. (2007). Maximum patterns in datasets. *Discrete Applied Mathematics 156*, 846-861.

Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A. (1997). Logical analysis of numerical data. *Mathematical Programming, 79*, 163-190.

Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE transactions on knowledge and data engineering, 12* (2), 292-306.

Colton, S. (2004). *Decision tree learning.*Imperial College London. Consulté en 2009, tiré de http://www.doc.ic.ac.uk/~sgc/teaching/v231/lecture11.html

Crama, Y., Hammer, P. L., Ibaraki, T. (1988). Cause-effect relationships and partially defined Boolean functions. *Annals of Operations Research 16,* 229-326.

Gorin, T., Brunger, W. G., White, M. (2006). No-show forecasting: A blended cost-based PNR-adjusted approach. *Journal of Revenue and Pricing Management 5* (3), 188-206.

Hammer, A. B., Hammer, P. L., Muchnik, I. (1999). Logical analysis of Chinese labor productivity patterns. *Annals of Operations Research 87*, 165-176.

Hammer, P. L., Kogan, A., Simeone, B., Szedmak, S. (2004). Pareto-optimal patterns in logical analysis of data. *Discrete Applied Mathematics 144* (1-2), 79-102.

Hammer, P. L., Bonates, T. O. (2006). Logical analysis of data – An overview: From combinatorial optimization to medical applications. *Annals of Operations Research 148*, 203-225.

Hillier F. S. (1998). A tutorial on optimization in the context of perishable-asset revenue management problems for the airline industry. In Saigal, R., Nagurney, A., Zhang, D., Padberg, M., Rijal, M., Vanderbei, R., Jaiswal, N., Gal, T., Greenberg, H., Prabhu, N., Fang, S. C., Rajasekera, J., Tsao, H., YU, G (éd.), Operations Research in the Airline Industry (pp. 68-98). Boston: Kluwer Academic Publishers.

Hipp, U., Nakhaeizadeh, G. (2000). Algorithms for Association Rule Mining – A general survey and Comparison. *SIGKKD Explorations 2* (1), 58-64.

Lawrence, R. D., Hong, S. J., Cherrier, J. (2003). Passenger-based predictive modeling of airline no-show rates. *SIGKDD 3*, 24-27.

Matteucci, M. (2008). *A tutorial on Clustering Algorithms.* Politecnico de Milano. Consulté en 2009, tiré de http://home.dei.polimi.it/matteucc/Clustering/tutorial_html

Mayoraz, E. (1995). C++ tools for logical analysis of data (RTR-1-95). New Jersey: Rutgers University's Center for Operations Research.

ANNEXE 1 – Guide de l'utilisateur

LAD-Datascope 2.0 Standard Edition, 2002

User Manual

Author: Sorin Alexe

RUTCOR, Rutgers Center for Operations Research

salexe@rutcor.rutgers.edu

http://rutcor.rutgers.edu/~salexe

Legend:
Commands, Buttons, Menu Items
Forms, Dialogs, Spreadsheets
Parameters, Options

1. Data

Data can be prepared using the **Observations** form (Menu: **LAD/Data**, or speedbutton **Dataset**). The **Observations** form contains a workbook with three spreadsheets: **Observations**, **Training** and **Test**. Each observation is represented as a row, first column containing the class (0 or 1), and all its attribute values (real numbers, or binary codes) in the next columns. Missing data are allowed, but not in the first row. Each of the spreadsheets can be edited, copy/paste data transfer is available through Workbook Designer (right double-click). The Observations form provides three different method for the construction of the **Training** and **Test** spreadsheets:

- **Random TRA/TES** - generates a stratified random partition,
- **k-folding TRA/TES** - divides the dataset into k stratified subsets, one of which is used as the test set, and all the others as training set. It is recommended to **Mix up** data before using this sampling procedure,
- **Leave one out** - the test set contains the observation specified by the **Index to test**, everything else is considered as training.

The LAD model is constructed on the basis of the training set only. The test set is used only for the evaluation of the inference power of the classification model.

Optionally, labels for attributes and observations can be added, attributes may be eliminated, or graphic representations of data/attributes can be visualized.

2. Discretization

LAD transforms the dataset into a discrete one by constructing a system of cutpoints. Thus, each value of an attribute is replaced by the index of that interval (in the list of intervals generated by the cutpoints corresponding to that attribute) which contains it. The system of cutpoints can be generated using the following options:

- **Bins** - construct the cutpoint system partitioning the dataset in almost equal sets, with respect to each individual attribute - unsupervised procedure,
- **Intervals** - construct the cutpoint system partitioning the dataset in almost equal intervals, with respect to each individual attribute - unsupervised procedure,
- **Cluster-based** - for each attribute, cutpoints are selected such that they separate the k-mean clusters of the values of the attribute,
- **Separation** - from the collection of all candidates, a separating set is selected.

To create a **Bins, Intervals** or **Cluster-based** cutpoint system:

- Define the **Default Grid Resolution** and
- Press **Set # cutpts,**
- **OPTIONAL:** change the number of cutpoints to be generated, spreadsheet/column **# of cutpoints,**
- Press **Make Cutpoints.**

To create a **Separation** cutpoint system:

- Define the **Separation [L]**evel - the program tries to separate any positive observation from any negative one by at least so many cutpoints,
- **OPTIONAL:** define **Limited [L]**evel, the program cannot produce more cutpoints for each individual attribute,
- Press **Make Cutpoints.**

To edit/print the cutpoint system, or to change the number of decimals, view **Grid Report**. Starting with column 2, for each attribute the corresponding cutpoints have to be listed (without gaps) in increasing order in the form **Cutpoints.**

To define a cutpoint system that is minimal, with equal number of cutpoints for all attributes, press **Min Eq Supp.** To find an minimal support set, press **Min Eq Supp** and **Min'l Supp.**

To increase the grid resolution use **Increase** (**By**: adds a number of cutpoints / **Times**: multiply by a number the column **# of cutpoints**). Press **Make Cutpoints**, again**.**

Discrete Space and a **Visualize** form for the visualization of the datapoints and of the grid, projected on two dimensions, are available, too.

To check the quality of a grid, press **Check Grid Quality.** If the grid separates the datapoints, the entropy equals zero, and separability indicates the minimum number of cutpoints that separate positive observations from negative ones. If the grid does not separate the datapoints, a lower entropy indicates a more informative grid.

3. Patterns

Patterns can be produced either as **Prime Cones** or as **Prime Intervals.** The control parameters for pattern generation procedure are:

- **Homogeneity - positive -** the proportion of positive observations, out of those which are covered, has to be higher than this parameter in order to keep the interval as positive pattern; **- negative -** the proportion of positive observations has to be lower than this parameter in order to keep the interval as negative pattern,
- **Prevalence -** indicates the lower bound for the proportion of observations in the corresponding class to be covered,
- **Degree -** the maximum number of attributes allowed for the description of the patterns,
- **# of Pat's** - if the number of produced patterns is larger than this parameter the pattern generation procedure is interrupted.

Patterns are filter with respect to their redundancy. If a pattern is included in another one (of the same type), it is deleted from the final collection.

The **Strongness** command will keep only one pattern out of those patterns covering exactly the same set of observations (in the class).

If patterns are added from outside the software, before any other processing they have to be re-**Evaluate**d.

Optional filtering can be performed after the corresponding new values of the control parameters are changed. This filter can be applied for **TRA**ining or for **TES**t.

4. Models

The program can extract from the collection of all available patterns (*the* pandect) a minimal sub-collection having the same explanatory power, i.e., all observations which are covered by some patterns in the original collection remain covered by some pattern in the smaller collection. The type of coverage can be controlled by the **Cover (C)** parameters, showing how many times has to be covered each observation. Pandect based classification uses all pattern to construct the *prognostic index* (a margin classifier). The prognostic index may use the **Equal Weights** or the **Prevalence Weights** to weight the patterns. While the first 4 options construct also the *pattern space*, i.e., the observations-patterns incidence matrix - , the last one - **Pandect (no Pattern Space)** - does not construct this matrix.

To construct the model, press **Make Theory.**

Various reports are available:

- **Forecasting Report** - contains 3 spreadsheets: the 1st for the training set, the 2nd for then test set, and the 3rd for statistics (2x3 tables)
- **Pattern Report** - contains 2 spreadsheets: 1st for statistics (2x4 tables), and the 2nd for model description: each row describes a pattern as the conjunction of conditions; a condition is described by the corresponding entry and it is imposed on the attribute in the corresponding column.
- **Feature Analysis** - is a special form that assigns to each attribute the degree of involvement in the description of the patterns.
- **Pattern Space** - visualize the pattern space as a matrix, or as a picture.

5. License

LAD-Datascope V2.0 LE (Light Edition) is a free software. It has 3 levels of access:

1. Demo
2. Limited (number of program launches)

3. Unlimited

The **Demo** access level does not require registering the software. There are several limitations on the size of the training sets to be used, as well as for the type of results that can be obtained. The **Limited** type of access allows the user to access all the features of the program. After the number of execution is achieved, the type of access is automatically changed to **Demo.** The **Unlimited** type of access allows the user to access all the features of the program, and never expire.

To obtain an access code, please read the agreement in **Registration form**, and send an e-mail to

Sorin Alexe: salexe@rutcor.rutgers.edu

http://rutcor.rutgers.edu/~salexe

The software comes with no warranty. It is not the author responsibility for any errors it might contain, or other problems that might occur while using it. However, there are not known hidden errors.A granted license allows the user to install LAD-Datascope V2.0 LE on a single machine. No part of this software can be sold, decompiled, reproduced or changed unless a special agreement with the author is made.

ANNEXE 2 – Résultats par produit

TANGO

		Prévisions			Mesures de l'erreur LAD			Mesures de l'erreur PROS		Meilleur
Vol	Présences réelles	Prévision LAD	Prévision PROS		Err.^2	Err. Rel.		Err.^2	Err. Rel.	PROS/ LAD
1	11	27.712	27.688		279.288	151.9%		278.473	151.7%	P
2	25	26.076	26.617		1.159	4.3%		2.616	6.5%	L
3	17	17.695	18.556		0.483	4.1%		2.420	9.2%	L
4	43	41.350	45.815		2.722	3.8%		7.925	6.5%	L
5	45	42.583	45.683		5.840	5.4%		0.466	1.5%	P
6	63	60.937	61.216		4.256	3.3%		3.181	2.8%	P
7	47	47.647	48.616		0.419	1.4%		2.612	3.4%	L
8	35	32.596	34.835		5.782	6.9%		0.027	0.5%	P
9	50	48.879	45.353		1.257	2.2%		21.592	9.3%	L
10	59	61.464	61.506		6.073	4.2%		6.280	4.2%	L
11	39	36.421	32.048		6.650	6.6%		48.335	17.8%	L
12	23	23.433	24.211		0.188	1.9%		1.466	5.3%	L
13	46	46.992	48.924		0.984	2.2%		8.550	6.4%	L
14	80	76.891	80.348		9.668	3.9%		0.121	0.4%	P
15	42	40.297	37.728		2.899	4.1%		18.249	10.2%	L
16	81	79.600	83.269		1.960	1.7%		5.149	2.8%	L
	Somme				329.625	207.8%		407.460	238.5%	
	Moyenne				20.602	13.0%		25.466	14.9%	
	Écart type				69.039	37.1%		68.576	36.7%	

Nbre P :	5
Nbre L :	11
LAD :	68.75%

TANGO PLUS

		Prévisions			Mesures de l'erreur LAD			Mesures de l'erreur PROS		Meilleur
Vol	Présences réelles	Prévision LAD	Prévision PROS		Err.^2	Err. Rel.		Err.^2	Err. Rel.	PROS/ LAD
1	26	34.844	34.432		78.218	34.0%		71.094	32.4%	P
2	16	17.503	14.584		2.260	9.4%		2.005	8.8%	P
3	38	37.184	42.717		0.667	2.1%		22.252	12.4%	L
4	51	47.659	49.992		11.164	6.6%		1.017	2.0%	P
5	21	19.836	22.886		1.355	5.5%		3.556	9.0%	L
6	38	34.761	28.584		10.492	8.5%		88.658	24.8%	L
7	42	39.495	40.259		6.277	6.0%		3.032	4.1%	P
8	40	41.165	37.914		1.357	2.9%		4.352	5.2%	L
9	23	25.458	26.112		6.043	10.7%		9.682	13.5%	L
10	20	19.002	17.571		0.996	5.0%		5.900	12.1%	L
11	24	22.144	21.491		3.447	7.7%		6.293	10.5%	L
12	14	14.648	16.626		0.420	4.6%		6.897	18.8%	L
13	64	57.802	57.956		38.413	9.7%		36.536	9.4%	P
14	24	24.930	20.127		0.864	3.9%		15.002	16.1%	L
15	38	37.927	26.696		0.005	0.2%		127.770	29.7%	L
16	14	18.349	17.654		18.914	31.1%		13.349	26.1%	P
	Somme				180.891	147.9%		417.394	235.1%	
	Moyenne				11.306	9.2%		26.087	14.7%	
	Écart type				20.396	9.5%		37.279	9.2%	
									Nbre P	6
									Nbre L	10
									LAD:	62.50%

LATITUDE

		Prévisions			Mesures de l'erreur LAD			Mesures de l'erreur PROS		Meilleur
Vol	**Présences réelles**	**Prévision LAD**	**Prévision PROS**		**Err.^2**	**Err. Rel.**		**Err.^2**	**Err. Rel.**	**PROS/ LAD**
1	4	4.844	10.286		0.713	21.1%		39.510	157.1%	L
2	3	2.956	4.333		0.002	1.5%		1.778	44.4%	L
3	2	2.217	0.000		0.047	10.9%		4.000	100.0%	L
4	3	2.422	3.467		0.334	19.3%		0.218	15.6%	P
5	6	5.379	12.600		0.386	10.4%		43.560	110.0%	L
6	8	14.862	24.700		47.088	85.8%		278.890	208.8%	L
7	6	6.322	15.333		0.104	5.4%		87.111	155.6%	L
8	4	6.652	7.611		7.033	66.3%		13.040	90.3%	L
9	9	8.495	0.000		0.255	5.6%		81.000	100.0%	L
10	7	5.174	11.667		3.335	26.1%		21.778	66.7%	L
11	2	2.217	4.000		0.047	10.9%		4.000	100.0%	L
12	1	7.437	22.750		41.436	643.7%		473.063	2175.0%	L
13	2	2.422	4.286		0.178	21.1%		5.224	114.3%	L
14	4	2.956	7.600		1.089	26.1%		12.960	90.0%	L
15	6	4.836	11.063		1.356	19.4%		25.629	84.4%	L
16	6	4.631	10.136		1.874	22.8%		17.110	68.9%	L
	Somme				105.278	996.2%		1108.870	3681.0%	
	Moyenne				6.580	62.3%		69.304	230.1%	
	Écart type				14.854	156.6%		127.580	520.7%	

Nbre P	1
Nbre L	15
LAD:	93.75%

AFFAIRES

		Prévisions			Mesures de l'erreur LAD			Mesures de l'erreur PROS		Meilleur
Vol	**Présences réelles**	**Prévision LAD**	**Prévision PROS**		**Err.^2**	**Err. Rel.**		**Err.^2**	**Err. Rel.**	**PROS/ LAD**
1	0.000	0.000	0.000		0.000			0.000		L
2	2.000	1.768	1.914		0.054	11.6%		0.007	4.3%	P
3	5.000	4.218	6.500		0.612	15.7%		2.250	30.0%	L
4	4.000	3.401	3.800		0.359	15.0%		0.040	5.0%	P
5	3.000	2.856	5.500		0.021	4.8%		6.250	83.3%	L
6	0.000	0.000	0.000		0.000			0.000		L
7	5.000	4.353	7.016		0.419	12.9%		4.064	40.3%	L
8	4.000	2.585	3.400		2.003	35.4%		0.360	15.0%	P
9	2.000	1.904	1.600		0.009	4.8%		0.160	20.0%	L
10	3.000	3.222	1.100		0.049	7.4%		3.610	63.3%	L
11	1.000	0.952	0.250		0.002	4.8%		0.563	75.0%	L
12	0.000	2.449	2.250		5.999			5.063		P
13	7.000	2.585	6.880		19.495	63.1%		0.014	1.7%	P
14	2.000	2.816	14.667		0.666	40.8%		160.444	633.3%	L
15	2.000	1.633	1.848		0.135	18.4%		0.023	7.6%	P
16	4.000	3.672	4.036		0.108	8.2%		0.001	0.9%	P
	Somme				29.930	242.8%		182.849	979.8%	
	Moyenne				1.871	18.7%		11.428	75.4%	
	Écart type				4.934	17.5%		39.793	170.1%	

Nbre P	7
Nbre L	9
LAD:	56.25%

AÉROPLAN

		Prévisions			Mesures de l'erreur LAD			Mesures de l'erreur PROS		Meilleur
Vol	**Présences réelles**	**Prévision LAD**	**Prévision PROS**		**Err.^2**	**Err. Rel.**		**Err.^2**	**Err. Rel.**	**PROS/ LAD**
1	6	8.516	9.000		6.330	41.9%		9.000	50.0%	L
2	12	11.405	11.754		0.354	5.0%		0.061	2.1%	P
3	7	6.951	7.810		0.002	0.7%		0.655	11.6%	L
4	4	5.433	5.000		2.053	35.8%		1.000	25.0%	P
5	8	6.804	6.535		1.430	14.9%		2.145	18.3%	L
6	7	5.434	6.286		2.452	22.4%		0.510	10.2%	P
7	3	5.140	3.939		4.579	71.3%		0.882	31.3%	P
8	6	5.286	5.273		0.509	11.9%		0.529	12.1%	L
9	21	21.240	22.222		0.058	1.1%		1.494	5.8%	L
10	23	22.073	21.125		0.860	4.0%		3.516	8.2%	L
11	4	3.769	3.060		0.054	5.8%		0.884	23.5%	L
12	4	3.769	3.524		0.054	5.8%		0.227	11.9%	L
13	10	9.495	9.714		0.255	5.1%		0.082	2.9%	P
14	8	7.244	7.217		0.572	9.5%		0.614	9.8%	L
15	3	3.329	2.674		0.108	11.0%		0.106	10.9%	P
16	4	3.329	3.652		0.451	16.8%		0.121	8.7%	P
	Somme				20.120	262.9%		21.826	242.2%	
	Moyenne				1.258	16.4%		1.364	15.1%	
	Écart type				1.821	18.8%		2.227	12.3%	

Nbre P	7
Nbre L	9
LAD:	56.25%